DIA data mining in colorectal cancer research

Dissertation

Zur Erlangung des akademischen Grades

Doctor rerum naturalium

Dr. rer. nat.

an der Fakultät für Mathematik, Informatik und Naturwissenschaften
am Fachbereich Chemie der Universität Hamburg

vorgelegt von

Oliver Kardell

Hamburg, August 2020

1. Gutachter: Prof. Dr. Hartmut Schlüter

2. Gutachter: Prof. Dr. Chris Meier

Tag der Disputation: 6.11.2020

Diese Arbeit wurde in der Zeit von April 2017 bis einschließlich Juli 2020 am Universitätsklinikum Hamburg-Eppendorf, Institut für Klinische Chemie, im Arbeitskreis massenspektrometrische Proteomanalytik unter Anleitung von Herrn Prof. Dr. Hartmut Schlüter angefertigt.

I. Table of Contents

II. Abbreviations

Name	Abbreviation
Acetonitrile	ACN
Chromosomal instability	CIN
Coefficient of variation	CV
Colorectal cancer	CRC
Common internal retention time peptides	CiRTs
Data-dependent acquisition	DDA
Data-independent acquisition	DIA
Dithiothreitol	DTT
dotProduct	dotP
False discovery rate	FDR
Fecal occult blood test	FOBT
Formic acid	FA
High pH	HpH
Indexed retention time peptides	iRTs
Iodoacetamide	IAA
Liquid chromatography	LC
Mass spectrometry	MS
Microsatellite instability	MSI
Multiple reaction monitoring	MRM
Peptide query parameters	PQP
Peptide-spectrum matches	PSM
Search engine combination of Comet and MS-GF+	CM
Search engine combination of Comet and X!Tandem	CT
Search engine combination of Comet, MS-GF+, and X!Tandem	CMT
Search engine combination of MS-GF+ and X!Tandem	MT
Search engine Comet	C
Search engine MS-GF+	M
Search engine X!Tandem	T
Selected reaction monitoring	SRM
Sequential windowed acquisition of all theoretical mass spectra	SWATH
Transfer of identification confidence	TRIC

1. Zusammenfassung

Darmkrebs stellt den am zweithäufigsten diagnostizierten Krebs dar und ist damit ein Hauptgrund für krebsverursachte Tode in der Welt. Bemerkenswert ist, dass sich in den vergangenen Jahrzehnten die Überlebensrate kaum geändert hat. Speziell in späteren Krebsstadien sinkt die 5-Jahres-Überlebensrate auf unter 10% [1,2]. Besonders die Erforschung der Pathogenese von Darmkrebs ist essenziell, um Früherkennungstests sowie neue Therapieansätze zu entwickeln. Hierbei spielen Fortschritte auf dem Gebiet der Proteomik, der Erforschung der Proteinzusammensetzung einer Zelle, eine entscheidende Rolle [2,3]. Das Hauptziel der Dissertation umfasste die Erforschung des Proteoms von Darmkrebs in verschiedenen Stadien, um potenziell signifikante Proteine herauszustellen. Das dabei identifizierte Protein-Panel sollte als Grundlage dienen, um mögliche neue pathogene Muster von Darmkrebs aufzudecken.

Die ausgewählte Strategie für die Erkennung von darmkrebsassoziierten Proteinen basierte im Kern auf einer labelfreien LC-MS/MS Methode inklusive *data-independent acquisition* (DIA). Zunächst wurde eine bioinformatische Pipeline entwickelt, um die hohe Informationsdichte der DIA-generierten MS2-Spektren bestmöglich zu nutzen. Dabei wurden mehrere datenbank-basierte Suchmaschinen für die Interpretation von MS2-Spektren kombiniert und die Ergebnisse in den jeweiligen Bibliotheken für eine darauffolgende Analyse der DIA-generierten Daten zusammengefügt. Der Einfluss einzelner Suchmaschinen oder mehrerer kombinierter Suchmaschinen auf die Analyse der DIA-Spektren wurde hinsichtlich der Größe der Bibliothek, der Konsistenz in der Datenanalyse, der Quantifizierungsleistung sowie der Identifizierung statistisch relevanter Proteine bewertet. Darüber hinaus wurden die bioinformatische Analysezeit und der Datenspeicherplatzbedarf einzelner Daten-analyseabläufe verglichen und in eine Gesamtevaluierung miteinbezogen. Als Input für die entwickelte Proteomik-Pipeline wurden einerseits *data-dependent acquisition* (DDA) Messungen von Darmkrebsgewebeproben, die vorher mittels *HpH-reversed phase* fraktioniert wurden, gewählt. Dieser Ansatz wurde als „DDA-based“ bezeichnet. Andererseits wurden die DIA-Messungen direkt als Input für die bioinformatische Pipeline ohne die Verwendung von DDA-generierten Daten benutzt. Diese Strategie wurde „DDA-free“ genannt.

Der „DDA-based" Ansatz zeigte, dass die Identifikationsrate auf der Ebene der Bibliothek steigt, wenn mehrere Suchmaschinen kombiniert werden. Außerdem ging die Formation einer binären Kombination aus Suchmaschinen sowohl mit einem Anstieg der Analysezeit als auch mit einem erhöhten Bedarf an Speicherplatz einher. Des Weiteren wies die DIA-Analyse darauf hin, dass eine erhöhte Informationsdichte in der Bibliothek keine bessere Quantifizierung der DIA-Daten garantiert. Die Resultate deuteten an, dass insbesondere die Retentionszeit und die Qualität der Bibliothekseinträge hinsichtlich der Signalintensität von essenzieller Bedeutung sind. Zudem demonstrierte die statistische Evaluierung, dass es wesentliche Unterschiede bei der Identifizierung signifikanter Proteine gibt, wenn unterschiedliche Suchmaschinen oder Kombinationen an Suchmaschinen verwendet werden für den Datenanalyseprozess.

Ähnliche Resultate wurden bei der „DDA-free" Strategie erzielt. In den meisten Fällen stieg die Identifikationsrate bei der Verwendung mehrerer Suchmaschinen auf Bibliotheksebene an. Außerdem zog die Kombination mehrerer Suchmaschinen einen erheblichen Anstieg der Analysezeit und des Datenspeicherbedarfs nach sich. Zudem wurde gezeigt, dass es keinen proportionalen Zusammenhang zwischen der Informationsdichte der Bibliothek und der Sensitivität der DIA-Analyse gibt. Darüber hinaus wurde bestätigt, dass die Auswahl der Bibliothek einen zentralen Einfluss auf die Identifizierung signifikanter Proteine hat. Im Vergleich beider Ansätze schnitt die „DDA-based" Strategie hinsichtlich einer höheren Identifikationsrate auf Bibliotheksebene sowie bei der Analysezeit und dem Speicherbedarf besser ab. Im Gegensatz dazu erreichte die „DDA-free" Methode eine höhere Anzahl an Quantifizierungsergebnissen der DIA-Daten.

Die Untersuchung der biologischen Bedeutung wurde für diejenigen statistisch signifikanten Proteine durchgeführt, die in beiden Analysestrategien identifiziert wurden. Die Analyse biologischer Prozesse und Netzwerke wies Korrelationen verschiedener detektierter Proteine in Entzündungsprozessen, in der Immunabwehr sowie der Aufrechterhaltung des zellulären Redoxgleichgewichts auf. Eine darauffolgende Literaturrecherche offenbarte mehrere Verbindungen der identifizierten Proteine zu bereits publizierten Resultaten im Kontext von Darmkrebs. Insgesamt stellen diese ermittelten Proteine eine exzellente Ausgangslage dar, um in Folgestudien mögliche neue pathogene Mechanismen von Darmkrebs zu untersuchen.

2. Abstract

Deciphering pathogenic mechanisms of colorectal cancer (CRC) is essential for understanding the development and the progression of the malignancy, as well as to establish detection in early stages and possible treatments [1-3]. The main aim of the thesis was to highlight significant proteins and elucidate potential pathogenic patterns by comparing the protein profile of CRC samples in different stages. To elaborate, the thesis aimed at identifying a promising protein panel which can be used as a valuable starting point for further research to decipher the pathogenesis of CRC. Here, the method of choice for the detection of CRC-associated protein profiles was a label-free LC-MS/MS strategy with data-independent acquisition (DIA).

First, a bioinformatic analysis workflow was implemented to exploit the high information input of the acquired digital DIA maps. The developed proteomic pipeline combined the results of multiple search engines to construct the corresponding libraries and to examine the influence of each generated library on the extraction of the DIA data. Moreover, two different data inputs were used for the bioinformatic workflow and the corresponding results were compared: Pre-fractionated data-dependent acquisition (DDA) measurements for the so called “DDA-based” analysis workflow and the DIA data for the analysis strategy termed “DDA-free”.

The DDA-based data analysis workflow demonstrated that the library input was increased by combining the results of multiple search engines. Furthermore, adding the results of a search engine to form a binary combination enhanced both analysis time and storage size. Besides, the DIA analysis indicated that there is no direct correlation between the increase of the library and the SWATH quantification performance. Further investigation suggested that the quality of library input regarding signal intensity and the retention time variability of the transitions are key characteristics in DIA data extraction. In addition, statistical evaluation showed that no database search engine combination achieved the detection of all possible statistically significant proteins.

The DDA-free data analysis strategy displayed similar results. Firstly, it demonstrated that in most cases merging the findings of one search engine to another search tool increased the identification rate on library-level. Secondly, combining multiple search engines had a significant impact on the analysis time and storage size. Further analysis indicated that an enhanced library input is not necessarily proportional to an improved performance of the DIA

analysis. Results showed that especially the quality of the library input regarding signal intensity and the retention time variability of the transitions have a substantial impact on the SWATH quantification performance. In addition, no database search engine combination with its corresponding library was able to identify all possible statistically significant proteins. Hence, the results suggested that the choice of library has a crucial influence on the detection of statistically significant proteins.

A comparison of the two data analysis strategies demonstrated that the DDA-free strategy achieved a smaller library input in comparison to the DDA-based strategy. On the other hand, the DDA-free approach obtained a better SWATH quantification performance and identified more statistically significant proteins. These results indicated that the quality of the library input is more significant than the total number of entries. Furthermore, experimental and computational requirements varied tremendously between the two data analysis strategies. The DDA-free approach had higher computational demands and the DDA-based strategy included higher experimental costs.

Statistically significant proteins which were identified in both data analysis strategies were submitted to biological inference. The pathway and network analysis demonstrated enriched biological paths in inflammation processes, immune responses, and maintenance of the cellular redox environment. In addition, literature mining revealed that the detected proteins had a previously described correlation to CRC. As a conclusion, the applied method including the data analysis strategy led to the discovery of a promising protein panel which serves as a valuable starting point for further studies in the ongoing research area of CRC.

3. Introduction

3.1 Colorectal cancer

Colorectal cancer (CRC) is the second most frequently diagnosed cancer and a major cause of cancer-related deaths in the world. In the past decades CRC survival rates have barely changed. After the development of metastasis, the 5-year survival rate is less than 10%, whereas it increases up to 90%, if CRC is detected early [1,2].

The development and progress of CRC is classified into five stages (Fig. 1). First, an adenoma, a benign precursor lesion, is formed (stage 0). After progression to a localized colon carcinoma (stage I and II), a CRC lymph node metastasis (stage III) is developed resulting ultimately in a spread to distant organs (stage IV) [4,5].

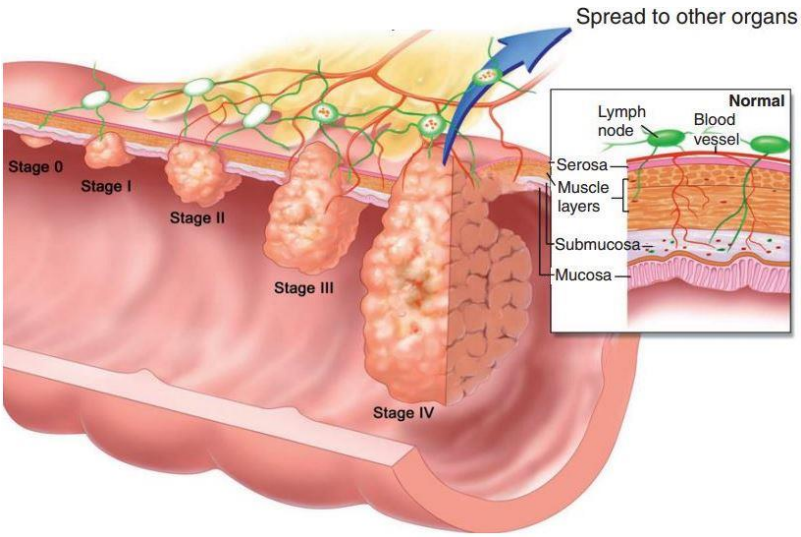


Fig. 1: Stages of colorectal cancer progression [5].

The process from a benign adenoma into cancer has an estimated duration of ten years and is often based on multiple genomic mutations. Frequent genetic alterations involve inactivation of tumor suppressor genes such as *TP53* or activating mutations in oncogenic pathways including *KRAS* and *BRAF*. The major causes for the genomic instability are the multiplication, deletion or translocation of whole chromosomes or of chromosome arms known as chromosomal instability (CIN). An additional reason is a defective DNA mismatch repair

machinery within nucleotide repeat sequences, called microsatellites, resulting in a so-called microsatellite instability (MSI). The vast difference in genetic alterations manifests itself in a heterogenic protein profile [4-6].

Up to date, the common clinically utilized screening strategies for early detection of CRC are the fecal occult blood test (FOBT) and colonoscopy. FOBT is successfully employed to reduce CRC mortality and is a simple, inexpensive, and non-invasive method. On the downside, it shows relatively low specificity, as well as poor sensitivity for the detection of CRC especially in early stages. Therefore, a follow-up detection by endoscope is often required. Colonoscopy presents a more reliable detection rate but is accompanied by inconvenience and invasiveness for the patient. Advances in genomics, the study of genes, or proteomics, the large-scale research of proteins, are the basis for further improving the understanding of pathogenesis of CRC and the development of new detection tests. The identification of genes or proteins that are characteristic for CRC are essential for progress in diagnosing CRC [2,3].

3.2 Role of biomarkers in CRC research

An important source for deciphering molecular mechanisms of CRC are biological markers or biomarkers [7]. The National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [8]. In general, biomarkers can be categorized into four groups – (1) diagnostic markers for early detection; (2) prognostic markers as indicators for the progression of the disease; (3) predictive markers for anticipating treatment response; (4) surveillance markers for monitoring disease recurrence [4,5]. In addition, the process from basic research via translational methods to clinical approval of biomarkers can be divided in a simplified way into three steps: discovery, verification, and validation (Fig. 2) [9].

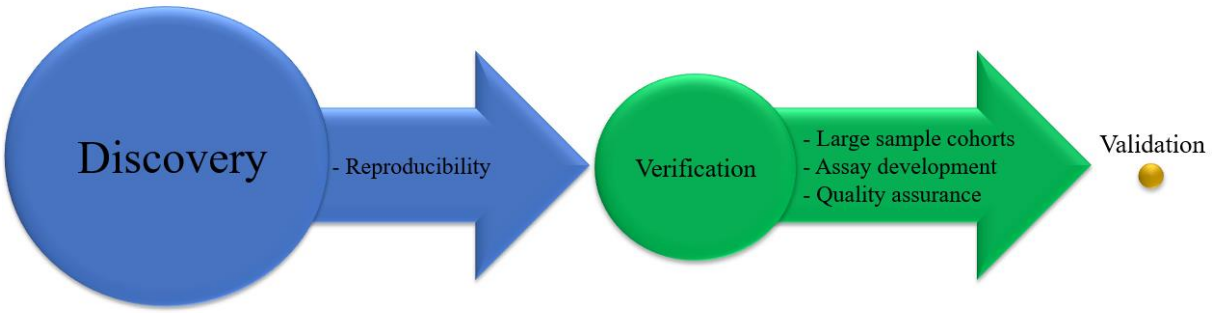


Fig. 2: From initial biomarker discovery into clinical implementation in three steps – discovery, verification, and validation. Circle sizes indicate proportion of outputs for the given step. Arrows contain major challenges to get to the next level.

While there is a plethora of methods for the discovery of potential biomarkers based on different high-throughput OMICs approaches, such as genomics or proteomics, measurement inconsistency and a lack of reproducibility across platforms and laboratories remain an obstacle in the verification process of the results [3,4]. Additional challenges arise before a successful validation. The availability and measurements of large sample cohorts including a high diagnostic accuracy, robustness of sample processing, and standardized data analysis procedures present a huge barrier for implementing validated clinical assays. The difficulties from initial discovery to a validated biomarker are reflected in an estimated success rate of 0.1% for clinical translation [3,4,9].

The clinical importance of CRC biomarkers in the context of pathogenesis and 5-year survival rate is shown in Fig. 3. Especially diagnostic biomarkers for the detection of progressive adenomas and early stage cancer have a significant clinical need, because the 5-year survival rate lies approximately at 90% [10,11].

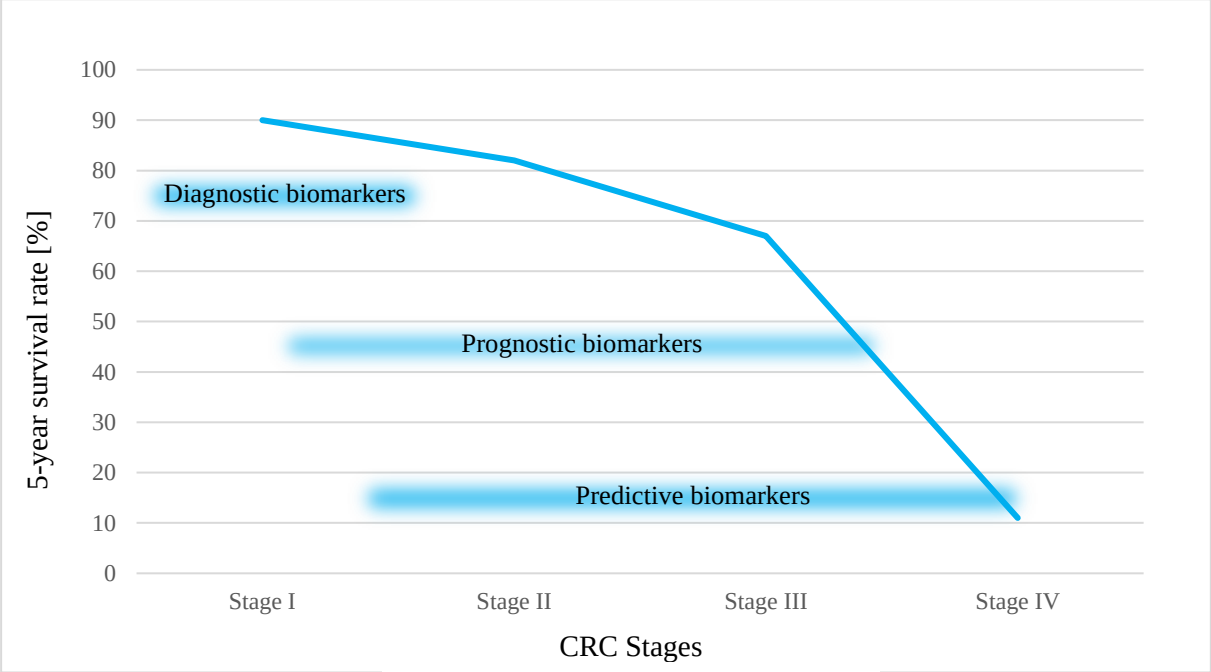


Fig. 3: Types of biomarkers in connection with CRC development and 5-year survival rate [%]. Figure adapted from Jimenez *et. al.* [10].

A prominent biomarker for CRC is the protein carcinoembryonic antigen (CEA). Originally considered specific for CRC, elevated CEA levels were also found in gastric and pancreatic cancers. Thus, the applicability of CEA as diagnostic marker for CRC is diminutive. However, it remains the marker of choice to monitor disease recurrence. Furthermore, determining the CEA level is well established in clinical routine work [12]. Potential molecular prognostic biomarkers include adenomatous polyposis coli (APC) and S100A2 protein. Mutation of APC is predicted as an indicator for the progression of CRC and high expression of S100A2 is correlated with tumor growth. Both biomarkers provide the potential of evaluating the development of the disease. Nonetheless, adequate validation is still necessary [7]. A possible predictive biomarker for CRC is the detection of *KRAS* mutations. Being part of epidermal growth factor receptor (EGFR) signaling pathways, discovering these mutations can be exploited to anticipate the response to anti-EGFR antibody-based therapies [7,12]. Recently, a promising diagnostic biomarker for early detection has been identified – microtubule associated protein RB/EB family member 1 (MAPRE1). In several studies MAPRE1 was differentially expressed in early neoplasm samples in comparison to healthy controls. Again, further validation of the findings is necessary emphasizing the discrepancy between discovery of potential biomarkers and validation with the final goal of broad clinical applicability [13].

However, many efforts have been made to improve the translational process [9]. In particular, multidisciplinary research approaches focus not on single biomarkers, but rather on biomarker panels by considering the vast heterogeneity of CRC. These strategies have the prospect of enhancing the sensitivity, specificity, and hence the diagnostic value of clinical assays. For example, a protein biomarker panel, which reflects the physiological state of the cell and the phenotype of the disease, in combination with genomics data has the potential to provide advancements in the translational process [9,13]. Moreover, expertise in bioinformatics big data analysis grows and highly reproducible pipelines are under development. In the future, further optimization in sample processing, detection methods, and computational strategies will gradually close the gap between initial biomarker discovery and a successful clinical translation to meet the need for CRC biomarkers [3,9,13].

3.3 Proteomics

Proteomics is primarily based on mass spectrometry (MS) and the corresponding methods have been widely applied to get new insights into biological mechanisms by deciphering the proteome and highlighting the role of proteins in cellular interaction networks and elucidate expression patterns in diseases [14,15].

In top-down proteomics the protein is investigated as an intact entity, which has the advantage of a detailed study and characterization of the molecular composition [16]. The different molecular forms at the genetic, transcriptional, or post-translational level of the same protein are called proteoform [17]. While maintaining the intramolecular complexity of the proteoform during analysis, common challenges arise from the lack of intact fractionation methods that are compatible with tandem MS. However, several efforts have been made to overcome limitations and difficulties to advance and exploit the potential of analyzing intact proteins [16-18].

In bottom-up proteomics, proteins are digested into peptides using trypsin or other proteases prior to the analysis via liquid chromatography and tandem MS [16]. The most widespread workflows can be categorized into three approaches – data-dependent, targeted, and data-independent proteomics (Fig. 4) [19].

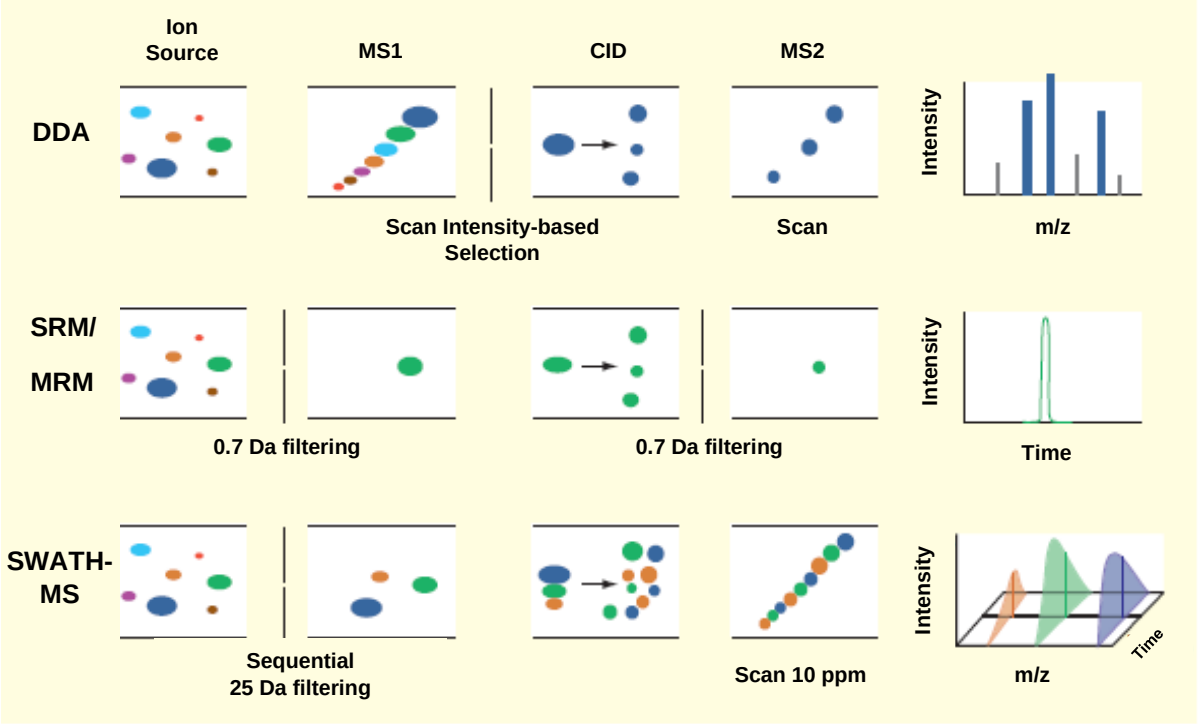


Fig. 4: Bottom-up proteomics – MS instrumental principles of DDA-based, targeted, and DIA-based proteomics [20].

Data-dependent acquisition (DDA) based proteomics is a universally and successfully applied approach with the goal of a complete coverage of the proteome identifying thousands of proteins in complex samples [19]. To elaborate, DDA involves a survey scan followed by the generation of MS/MS data. During the survey scan an automatic selection of precursor ions above a pre-set abundance threshold and fragmentation of the selected precursor ions takes place resulting in a MS/MS full scan on the product ions. So, the selection of which ions get fragmented is *dependent* upon some criteria previously set. In a typical LC-MS/MS experiment, the acquisition of tandem MS data is triggered by the precursor ion intensity. Over the course of the entire LC run, MS/MS data is generated from the most abundant precursors. Additional optimization is achieved by omitting the re-sampling of the same precursor ion via dynamic exclusion filtering [19,20]. A resulting drawback is that precursors within the excluded mass range, which are not previously selected but eluting at similar times, are not subjected to fragmentation. Moreover, the intensity-based selection of precursor ions follows heuristic principles. Consequently, the run-to-run reproducibility suffers. DDA-based proteomics is prone to irreproducible protein identification and quantification across large sample cohorts undermining the potential of achieving a great protein depth per run [19-21].

Targeted strategies, such as selected reaction monitoring (SRM), are widely performed in a clinical context because of an enhanced reproducibility and accuracy. SRM is usually performed on triple quadrupole (QQQ) instruments. Precursor ions of a specific peptide are selected in the first quadrupole (Q1). After fragmentation in the second quadrupole (Q2), a specific fragment ion from the target peptide is filtered in the third quadrupole (Q3) and guided to the detector. Precise quantification is based on chromatographic traces representing intensity profiles of the fragment ion signals over time [20,22,23]. The method always measures predetermined pairs of peptide precursors and corresponding fragment ions, which is termed as transition. Hence, for establishing targeted strategies prior knowledge about the protein of interest is necessary involving information about the precursor m/z ratio and product m/z ratio of proteotypic peptides. This targeted fashion achieves high reproducibility across large sample cohorts but comes at the cost of low proteome coverage because of a limited number of targetable proteins per MS injection. A typical application of SRM is restricted to the targeted measurement of up to 100 proteins per run [22,23].

Recent advances in MS instrumentation, which include new hybrid instruments like the quadrupole-TOF (Q-TOF) or the Q-Orbitrap set of instruments, gave rise to the development of data-independent acquisition (DIA) – a possibility to combine the advantages of DDA-based proteomics to detect a high number of analytes with the favorable dynamic range, sample throughput, and reproducibility of SRM (Fig. 5) [24-26].

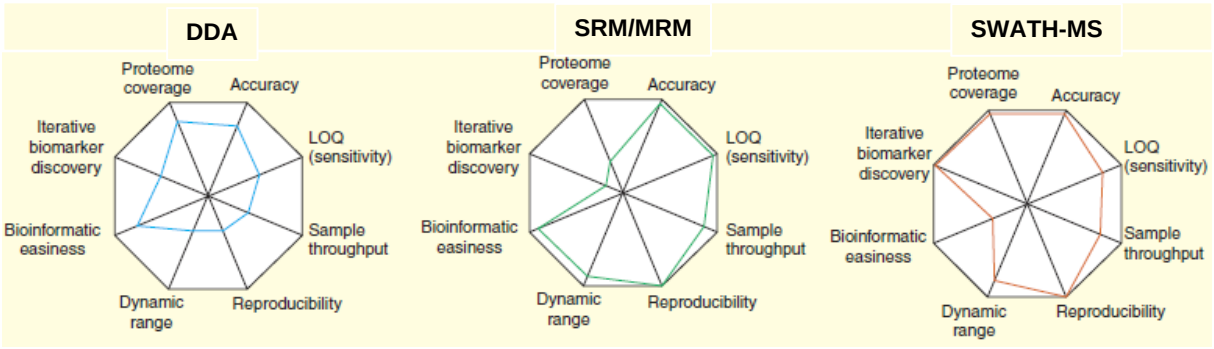


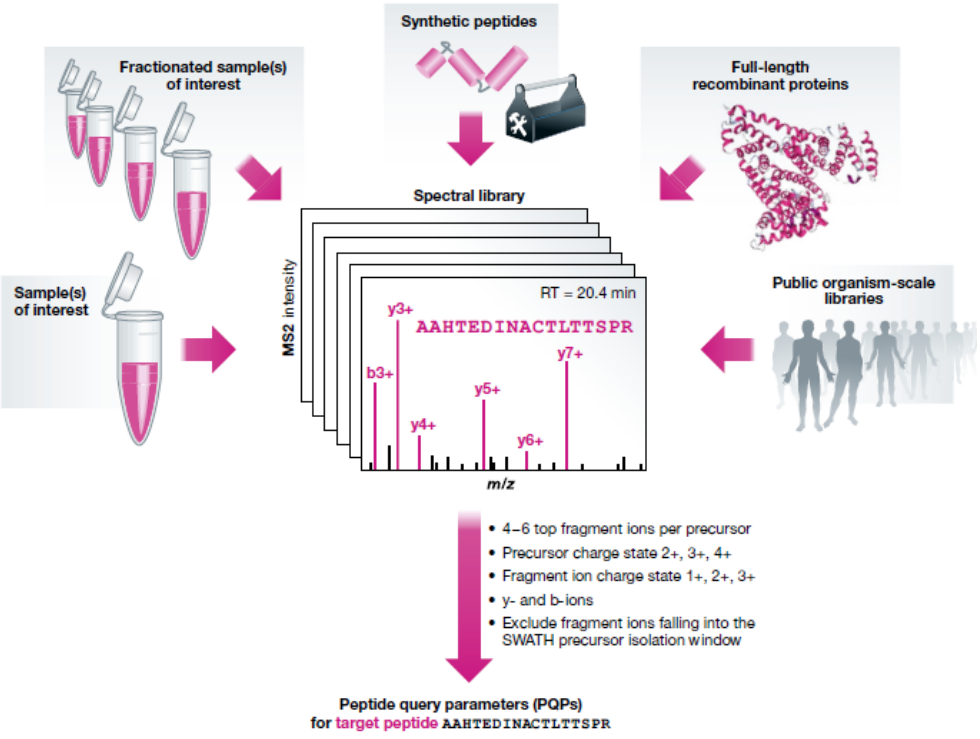
Fig. 5: Comparison of technical advantages and disadvantages of DDA, SRM, and SWATH-MS by performance profiles [20].

In DIA a fragmentation of all precursors at the same time within a certain m/z range is performed. In this case the fragmentation is *independent* of any characteristics of the precursor ions. As a result, complex tandem MS spectra are generated representing a digital map for the corresponding samples [20,27]. Deconvolution of the MS2 space and the complexity of data analysis remain bioinformatic challenges for DIA methods (Fig. 5) [20,28]. In contrast, for DDA-based proteomics and targeted approaches multiple pipelines and software tools are available and implemented in the proteomic community [29-31]. However, the promise of DIA to combine high reproducibility with great protein coverage drives research to the development and improvement of DIA-based methods [28].

3.4 SWATH-MS

All DIA-based methods rely on the same principle of continuously acquiring fragment ion spectra in an unbiased fashion [19]. Since the year 2000, in which Masselon *et al.* presented a proof of principle experiment for the simultaneous MS/MS analysis of multiple peptides and further development by Venable *et al.* in 2004 by the application of sequential precursor windows in tandem MS, several DIA strategies rest on the use of different types of mass spectrometer, distinct acquisition settings, and parameter optimizations, as well as analysis workflows [28,32,33]. Initially, DIA-generated data was directly submitted to DDA analysis tools due to a lack of specific software and data analysis pipelines for DIA data [34]. In 2012, Gillet *et al.* presented a new method called SWATH-MS, which combined unbiased DIA with a novel targeted data extraction strategy [27]. In this case, sequential windowed acquisition of all theoretical mass spectra (SWATH) is achieved by repeatedly cycling through 32 consecutive 25-Da precursor isolation windows resulting in a data set, which is continuous in retention time dimension and fragment ion intensity. The digital fragmentation ion maps are then mined by using information provided by a spectral library. The idea is that each peptide in the highly convoluted SWATH data can be uniquely identified by so called peptide query parameters (PQPs) in the spectral library. The peptide-specific information of the spectral library covers precursor and fragment ion signals, relative intensities, ion types and chromatographic parameters [27,28].

Since the development of SWATH-MS with targeted data extraction, many efforts have been made to ensure a high-quality and comprehensive library. For generation of prior knowledge and collecting the needed PQPs for the targeted data extraction several sample input types are utilized (Fig. 6) [28].



Protein	Peptide sequence	Precursor ion m/z	Fragment ion m/z	Precursor charge	Fragment charge	Fragment ion type	Fragment relative intensity	Normalized retention time
Raf1	AAHTEDINACTLTTSPR	929.4416	775.4308	2	1	y7	70	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	674.3832	2	1	y6	10	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	561.2991	2	1	y5	40	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	460.2514	2	1	y4	20	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	359.2037	2	1	y3	100	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	280.1404	2	1	b3	60	45.3

Fig. 6: Overview of input samples for generating a spectral library with peptide query parameters [28].

Usually DDA measurements of the same sample and on the same instrument are performed to acquire the PQPs. The coverage of single-shot DDA analysis is often lower than the corresponding SWATH-MS data. Therefore, repeated DDA analysis can be beneficial to increase sensitivity [24,28]. An additional approach for enhancing the information content of a spectral library is based on sample fractionation prior to DDA analysis [35,36]. Here, different fractionation strategies can be applied for further improvement [37]. Tandem MS spectra for library generation can also be derived from chemically synthesized peptides, which already

have proven their implementation as a valuable source of prior information e.g. in SRM assays [28]. Further extension of this idea has led to the development of synthetic full-length proteins by recombinant methods [38]. Another possible strategy relies on publicly available spectral libraries on an organism-scale [39]. In this context, important considerations about the transfer of information between instrument types and between laboratories, as well as appropriate global error rate control is required [20,39]. In principle, hybrid libraries of several approaches are also possible. In 2015 Schubert *et al.* generated a library consisting of endogenous samples and synthetic peptides [40]. The development and research regarding library generation for targeted data analysis is still ongoing, always optimizing for increased sensitivity and selectivity for an improved data mining of DIA measurements [28,34].

As a result of advancement in data analysis tools and technical improvements, proteomic researchers are able to perform SWATH-MS in a routine fashion to generate valuable biological insights [41]. In 2017, Yanzhang Luo *et al.* described the identification of carbonic anhydrase 2 (*CA2*) as a potential diagnostic biomarker for nasopharyngeal carcinoma by SWATH-based proteomics, which emphasizes the applicability of the DIA approach for clinical research [42]. However, to grasp a deeper understanding of potential challenges in SWATH-MS and the bioinformatic connection between DDA and DIA, a closer look on data analysis strategies and software tools in MS is beneficial [28,43].

3.5 Data analysis in bottom-up proteomics

In bottom-up proteomics the direct connectivity between proteins and experimental acquired spectra is lost. The proteins are digested by proteases into peptides, which are then analyzed via MS. The bioinformatic challenge is to reassemble peptides from the MS-based spectra and in a consecutive step to the related proteins. There are two basic strategies for the bioinformatic inference from the acquired spectra back to the protein: spectrum-centric and peptide-centric analysis (Fig. 7A; 7B) [44,45].

In spectrum-centric analysis, the query unit is a MS/MS spectrum. The approach assumes that each spectrum is generated from at least one peptide and the goal is to identify a peptide for

each spectrum, which best explains the data. The resulting assigned peptide-spectrum matches (PSMs) are subjected to statistical evaluation [44,45]. Especially, DDA measurements are analyzed by this concept [46]. On the other hand, peptide-centric analysis takes the peptides of interest as query units and looks for corresponding signals of each peptide in the MS/MS data. The underlying assumption is that each peptide elutes once during liquid chromatography. Statistical evaluation relies on the competition between candidate spectra from the acquired data for the best scoring evidence of detection. This approach is applicable for targeted strategies including SRM and DIA [44,45].

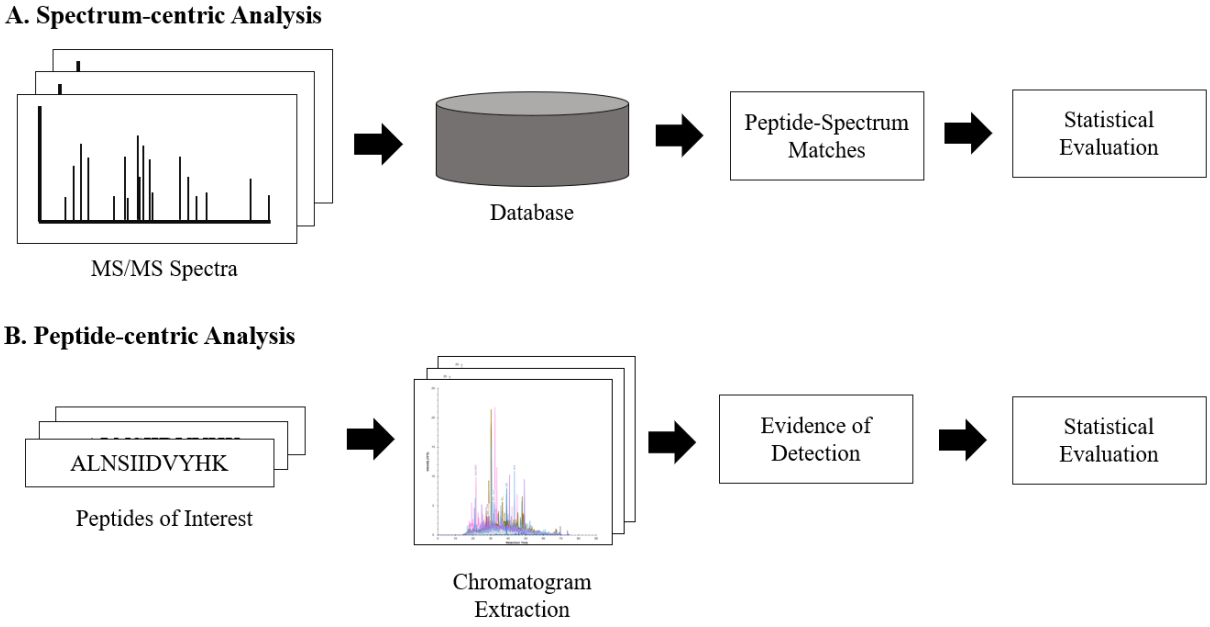


Fig. 7: Bioinformatic strategies for the analysis of tandem MS-data: spectrum-centric (A) and peptide-centric analysis (B) [44].

3.5.1 Data analysis in DDA

Large-scale shotgun proteomics is generally analyzed in a spectrum-centric manner [45]. After spectral processing of the raw MS data, the core element of the bioinformatic analysis is the peptide identification step. Acquired MS/MS spectra are interpreted by database searching, spectral library searching, or *de novo* sequencing. Statistical assessment and validation of peptide identification with the consecutive process of protein inference complete the data analysis workflow [47].

3.5.1.1 *De novo* sequencing

De novo spectrum identification is based on computationally inferring the sequence or partial sequence of peptides directly from the experimental tandem MS spectra by considering all possible amino acid combinations [48]. Hence, *de novo* methods avoid the necessity of a reference database, which makes it a powerful approach for the analysis of organisms with unsequenced or only partially sequenced genomes. On the downside, the computational expense is large and high-quality spectra are required for an effective analysis [47]. Nevertheless, great efforts have been made to establish *de novo* methods into daily data analysis routine for large scale proteomic data sets. Over twenty sequencing programs have been developed involving Lutefisk, PepNovo, and Twister [49-52].

3.5.1.2 Spectral library searching

Spectral library searching achieves peptide identification by comparison of the query MS/MS spectrum to a library of previously identified reference spectra [53]. The similarity of the spectra is mainly analyzed via a dot product scoring scheme [54]. A high-quality reference spectral library is a crucial prerequisite, because false positives can undermine the analysis [53]. Another drawback is that the peptide identification is limited to the content of the library [46]. However, spectral library search tools such as SpectraST and Bibliospec have the potential to exceed the performance of database search engines regarding speed and peptide identification rate [47,53,55].

3.5.1.3 Database searching

The dominant applied method for assigning peptides to tandem mass spectra is sequence database searching. Important steps contain preprocessing of the spectra, peptide identification, and error rate control [47].

3.5.1.3.1 Preprocessing

Spectral processing of the raw data has a direct impact of the peptide identification in terms of accuracy and specificity. The general goal is to detect and eliminate inconsistencies during MS acquisition [56]. Essential actions cover joining of redundant spectra, exclusion of low-quality spectra, and recognition of "chimeric" spectra, which are generated by two or more coeluting peptides. Applying and optimizing spectral processing steps can greatly enhance the outcome of peptide identification [47,56,57].

3.5.1.3.2 Peptide identification

All database search engines rely on the same principle. For a given spectrum *S*, a peptide database *P*, a precursor mass *m*, and a precursor mass tolerance δ, the algorithm defines candidate peptides *C*, which need to be part of the database *P* and the difference of the calculated mass *m(p)* of the peptide *p* and the corresponding precursor mass *m* has to be smaller than the precursor mass tolerance δ (Eq. 1) [46].

$$C(m, P, \delta) = \{p: p \in P; |m(p) - m| < \delta\} \quad \text{(Eq. 1)}$$

In the next step, a so-called scoring function Ξ(;) generates a theoretical fragmentation spectrum for each candidate peptide and compares it against the experimental acquired tandem mass spectrum. The likelihood of the peptide sequence match is scored and the peptide with the highest score is reported by the search tool (Eq. 2) [46,47].

$$\arg\max_{p \in C(m,P,\delta)} \Xi\,(S, p) \quad \text{(Eq. 2)}$$

3.5.1.3.3 Scoring functions

A multitude of database search engines has been developed, which mainly differ on the level of scoring function to infer theoretical spectra and to determine the degree of similarity between theoretical and experimental spectrum [58]. Primarily, scoring functions can be categorized into three different strategies: descriptive, interpretative, and stochastic modeling [59].

Descriptive models first use mechanistic predictions of fragmentation patterns of peptides for the generation of theoretical spectra and then assess the quality of a peptide spectrum match via a cross-correlation score [58,59]. Common database search engines such as SEQUEST, Comet, and X!Tandem achieve good sensitivity and applicability to different types of mass spectrometers and data sets [60-62].

Interpretative strategies infer the identification of peptides by extracting parts of the fragment ion series of a spectrum and using that partial sequence for the database search [59]. The extracted amino acid sequence is in the middle of masses of unknown composition, which gives the algorithm a broader flexibility [58]. Thus, powerful search tools such as TagRecon or MS-GF+ have been developed to identify mutations or to deal with the development of enhanced and novel MS techniques [63,64].

The stochastic approach relies on probability algorithms, which model theoretical spectra using training sets of spectra of known sequence identity [58]. The data mining process often utilizes machine learning algorithms for specific data sets offering the potential for an instrument tailored database search. A representative of a stochastic based search tool is SCOPE [59,65].

However, the assigned score of a database search tool for a PSM is either based on an arbitrary scale or converted to a statistical equivalent such as *p* value or expectation value. In each case further processing for statistical validation and an effective global error rate control for large-scale studies is required [47]. In addition, it is noteworthy that several studies have been performed, in which multiple search engines have been coupled to achieve a higher identification rate [66,67]. Especially beneficial to maximize the outcome of each proteomic dataset is to utilize search engines with distinct scoring function principles [66].

3.5.1.3.4 Postprocessing and protein inference

The foundation for reproducible results across platforms and datasets is an accurate error rate estimation. State-of-the-art tools for postprocessing scores of PSMs from different search tools include Percolator and PeptideProphet. Both convert search engines scores into probabilities and then compute a threshold to remove uncertain assignments [68-70]. The most common approach rests on estimating a false discovery rate (FDR), which is defined as the expected proportion of incorrect PSMs among all accepted PSMs [47]. Firstly, a global FDR is determined via a target-decoy database strategy, in which experimental tandem MS spectra are additionally searched against a database of proteins with reversed, randomized or shuffled sequences [47,71]. The number of matches from the decoy database presents an accurate estimate of false positives [68]. Secondly, a posterior probability for individual PSMs that estimate the correctness of the assignment is calculated and used to compute a baseline measure to differentiate between correct and incorrect identifications [47].

For protein inference, PSMs are grouped to their corresponding protein by performing additional evaluation of posterior probabilities and FDR estimation on protein-level [47]. Several programs have been developed to ensure an accurate transition from peptide-level FDR to protein-level error estimation, for example iProphet and MAYU [72,73]. Furthermore, there is a multitude of programs available for protein inference such as FIDO and ProteinProphet, which have been evaluated and benchmarked [74-76].

3.5.2 Data analysis in DIA

In DIA, MS/MS spectra are systematically acquired regardless of intensity. Parallel fragmentation of all detectable ions within a predefined m/z range generate complex digital maps of the MS2 space. Hence, to exploit the high information content of DIA data sophisticated data analysis workflows and software are essential [34,44]. Most dominant analysis strategies employ a peptide-centric approach with the use of a spectral library [43]. Analysis pipelines cover open-source software such as Skyline and OpenSWATH or commercially available programs like Spectronaut [24,77,78]. Additionally, the development

of spectrum-centric based software e.g. DIA-Umpire or Group-DIA, which do not necessarily rely on prior library generation, attract the attention of the research community [79-81].

3.5.2.1 SWATH-MS data analysis

A typical data analysis workflow for SWATH-MS data consists of library generation, DIA data extraction, probability assignment and validation, as well as quantification to infer statistical significance (Fig. 8) [28].

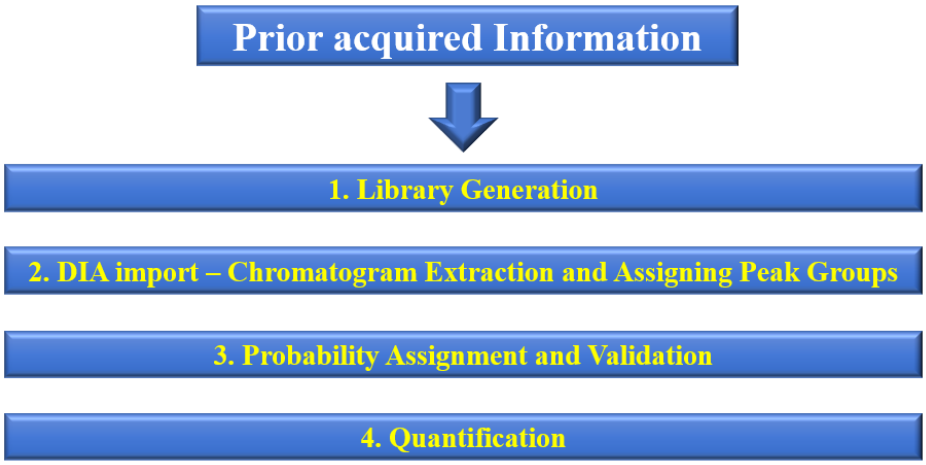


Fig. 8: Data analysis steps for SWATH-MS.

Quality and coverage of spectral reference libraries for peptide identification are of great value for targeted analysis [43]. While there is a multitude of possible input samples for assay libraries (see section 3.4), commonly DDA measurements performed under similar conditions and on the same instruments are employed assuring that the acquired MS/MS spectra resemble the relative fragment ion intensities in the SWATH-MS maps in a best possible way [28,43]. Moreover, several efforts have been made to improve the DIA extracting process including the optimization of retention time prediction with the use of indexed retention time peptides (iRTs) and further alignment via transfer of identification confidence for reproducible protein quantification (TRIC) [82,83]. However, the general fact, that only information about peptides, which are included in the library, can be used for DIA data extraction impacts considerations about both quality and coverage of spectral libraries. In terms of quality this stresses the

importance of reliable and accurate FDR control in order to avoid error propagation into DIA analysis [43]. Therefore, a bioinformatic link between DDA and DIA is based on the FDR control for library generation. Software like MAYU ensures an accurate estimation of the FDR on PSM-, peptide-, and protein-level in large-scale DDA data sets and thus is crucial so that only high-quality spectra with high-confidence peptide assignments enter the library [43,73]. For example, Bibliospec, which is implemented into Skyline, uses the cut-off score for a specified FDR reported by MAYU for library generation [43,55]. Regarding coverage of a spectral library, studies have shown that prefractionation prior to DDA measurements can enhance the information input in a subsequent library for DIA analysis [36,37]. Furthermore, it is recommended to use multiple search engines in an orthogonal way to increase the peptide identification rate. In summary, if DDA measurements are applied for library generation, DIA analysis cannot be implemented without making considerations about DDA analysis. Both coverage and quality of a library is greatly influenced on DDA level [43].

After library generation, chromatogram extraction of the DIA data including assigning peak groups and validation are the next steps in targeted SWATH-MS analysis [34]. First, precursor and fragment ion chromatograms for the peptides of interest are extracted with information of predefined PQPs stored in the library. In this context, the selectivity of extraction is influenced in retention time and mass tolerance dimension. A retention time window centered around the expected elution time is chosen with the aim of reducing the size as much as possible to enhance the accuracy of the extraction process. In addition, the width of ion extraction directly impacts selectivity of the chromatographic elution profile and thus optimization of the extraction width during the data analysis process is beneficial to increase identification rate and to improve peptide-centric scoring in a subsequent validation step [28]. Commonly, probability assignment of peak groups relies on a target-decoy approach. After generating decoy peptides with a reversed, shuffled or randomized sequence, fragment ion chromatograms are extracted next to target peptides [34]. Programs like mProphet, which is implemented into Skyline, or PyProphet, which is available in the OpenSWATH environment, calculate for target and decoy peptides several chromatogram- and spectrum-based scores [84,85]. All scores are combined into a discriminant score by a super-vised learning strategy. Subsequently, the distribution of peptide precursor count and corresponding discriminant score for both target and decoy peptides are used for FDR calculation [28,34]. In addition, Skyline has implemented another score to access

the similarity of peptide fragmentation patterns, which is referred to as dot product (dotP) and is based on a geometrical distance measure including a normalized spectral contrast angle [86].

Thorough peptide and protein quantification is the last step of SWATH-MS analysis [79]. Statistical strategies cover basic data processing steps, statistical modeling and inference of protein abundance [28]. First, peak intensities of peptides are transformed e.g. by normalization in order to prevent inter-run variation [87]. Recently, Narasimhan *et. al* studied and stressed the importance of the impact of normalization methods in SWATH-MS data analysis [88]. Next, the peak intensities are summed or averaged to infer the protein abundance of correlating peptides [87]. Again, several strategies and software have been developed [34]. For example, MSstats, which is integrated as Add-on in Skyline, employs a family of linear-mixed models for relative quantification of proteins and peptides [87]. Another option is Perseus, which offers several statistical tools for analyzing OMICs data including normalization, pattern recognition, as well as multiple hypothesis testing [89].

3.5.2.2 Alternative data analysis strategies

Spectrum-centric strategies such as DIA-Umpire or Group-DIA combine information of precursor and fragment ions of DIA data to generate pseudo-MS/MS spectra, which can be searched by conventional database search tools. Hence, prior to DIA analysis neither additional DDA measurements nor sample amounts are needed to generate a library [80,81]. DIA-Umpire performs a signal processing algorithm, which aims at detecting all possible MS1 peptide precursor ions and MS2 fragment ions (Fig. 9). For each monoisotopic peak of a precursor and fragment peak a Pearson correlation is calculated to build precursor-fragment groups. These co-eluting precursor and fragment ions form pseudo-tandem MS/MS spectra [80].

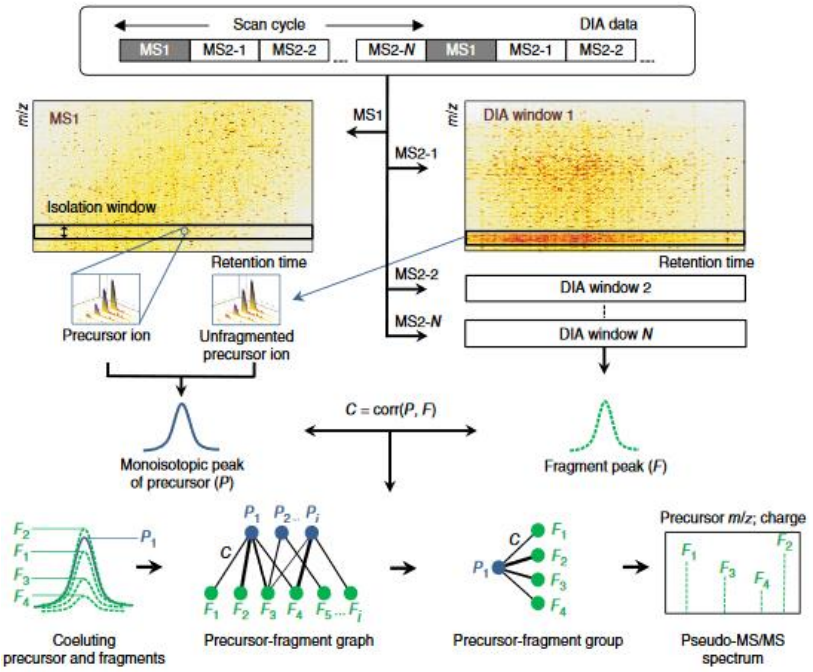


Fig. 9: DIA-Umpire signal processing to generate pseudo-MS/MS spectra directly from DIA data [80].

A recent study has demonstrated the applicability of spectrum-centric approaches for DIA data analysis and compared the performance to other widely used software methods. Results show that while peptide-centric analysis workflows exceed spectrum-centric strategies for low-quality spectra, similar results are achieved for high-quality spectra [90]. Thus, as instrument performance and corresponding measurement selectivity and sensitivity improve, generation of prior knowledge via DDA measurements for DIA analysis might become less significant [28].

4. Aim of the Thesis

Colorectal cancer (CRC) remains a major cause of cancer-related deaths in the world and in the past decades CRC survival rates have barely changed. Elucidating the development of CRC on molecular level from a benign precursor lesion in stage I to tumor metastasis in stage IV is of utmost importance [1,2]. In addition, the rise of new MS-based strategies, especially data-independent methods, which combine the potential of a great protein depth and outstanding consistency across large sample cohorts, open up new opportunities to decipher pathological patterns [20,79].

The main goal of the thesis was to highlight significant proteins and elucidate potential biological patterns in the sense of systems biology by comparing the protein profile of CRC samples in different stages. Hence, the thesis aimed at identifying a promising protein panel which can be used as a valuable starting point for further research to decipher the pathogenesis of CRC. The approach was based on the hypothesis that detection and classification of CRC in the future will be much more precise if the diagnostic target is not limited to one single protein but to a protein panel containing many proteins. Furthermore, the assumption included that despite a vast cancer heterogeneity different individual CRC will always have some proteins in common. All individual cancer cells must share a special inventory of proteins to survive in a healthy environment, which attacks them. The given hypothesis was based on several studies regarding cancer research published by the Schlüter group [91-93].

5. Workflow

The method of choice for a detection of CRC-associated protein profiles was a label-free LC-MS/MS strategy with data-independent acquisition (DIA). Overall, the strategy focused on creating a bioinformatic analysis workflow to exploit the high information input of the acquired digital DIA maps. To elaborate, the idea aimed at boosting the identification rate on library-level and subsequently the sensitivity of the DIA analysis by combining pre-fractionated DDA measurements with a data analysis including multiple search engines. Moreover, an approach included alternative strategies for DIA data analysis, which only require DIA measurements, to investigate potential merits in comparison with the first approach.

Consequently, the PhD thesis was divided into five main parts (Fig. 10). The first part was the development of an analysis workflow for library-based DIA data mining including a script for combining multiple search engines for peptide identification. The second part was based on using DDA spectra as input for the developed analysis workflow and depicted as “DDA-based” analysis. The benefits of using multiple search engine combinations were evaluated by several criteria such as analysis time, storage size, library size, and extraction of statistically significant hits. In the third part the same analysis procedure and evaluating scheme was performed only taking pseudo-MS/MS data directly generated by the DIA data without the need of DDA measurements, represented as “DDA-free” analysis. Importantly, the developed bioinformatic workflow was in its main structure applicable for both strategies in order to be as consistent as possible to ensure comparability. In part four both ways were compared and potential advantages and disadvantages discussed. Lastly, biological inference of potential significant proteins regarding CRC were addressed.

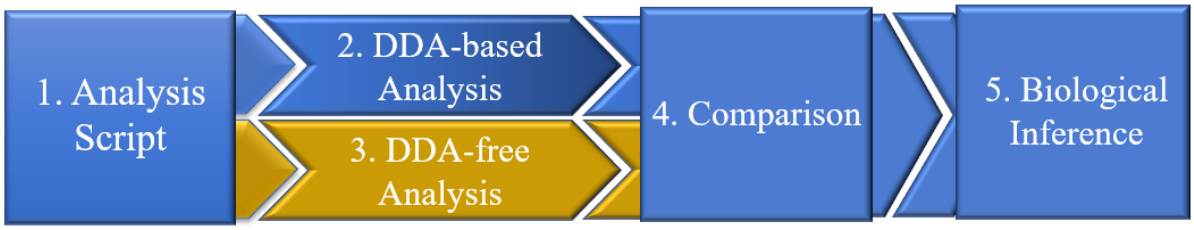


Fig. 10: Workflow of the PhD thesis – the developed analysis script in step one builds the framework for both DDA-based and DDA-free analysis prior to evaluation and comparison of both approaches in step four, as well as biological inference of significant patterns for CRC in step five.

6. Part I - Development of a DIA data analysis workflow

The foundation of the PhD thesis was the development of a DIA data analysis workflow, which guaranteed a high degree of reproducibility and the possibility of reiteratively processing the data. In addition, it needed to provide a certain flexibility in terms of using DDA spectra or pseudo-MS/MS spectra for a DDA-based or DDA-free analysis, respectively.

In general, the data analysis for the library-based approach can be divided in a simplified way into three main processes (see details starting from section 3.5.2):

1. Peptide Identification – generating prior knowledge
2. Library Generation
3. DIA Analysis

For peptide identification an adjustable, automated user-specified batch script was created. Library generation and DIA analysis were performed with Skyline and further statistical analysis was employed with the statistical software R. To understand possible benefits of using multiple search engines on the peptide identification level and to illustrate the influence on potential significant hits after statistical analysis on DIA level, it is crucial to elucidate the different parts of the data analysis workflow in detail. Therefore, in the following chapters the information flow will be described and important considerations will be highlighted.

6.1 Peptide identification – generating prior knowledge

The peptide identification covered preprocessing of the data, database search, and validation. In essence, the script combined programs for individual steps in a consecutive manner. Only open source software was implemented for the developed, automated analysis script for peptide identification. Integrating open-source software into a self-designed script had several merits such as transparency, repeatability, and adjustability. Especially the flexibility while maintaining a constant frame for data analysis was essential to perform and compare the DDA-based and DDA-free approach. The script for peptide identification for both DDA-based analysis and DDA-free analysis is shown in Fig. 11A and 11B.

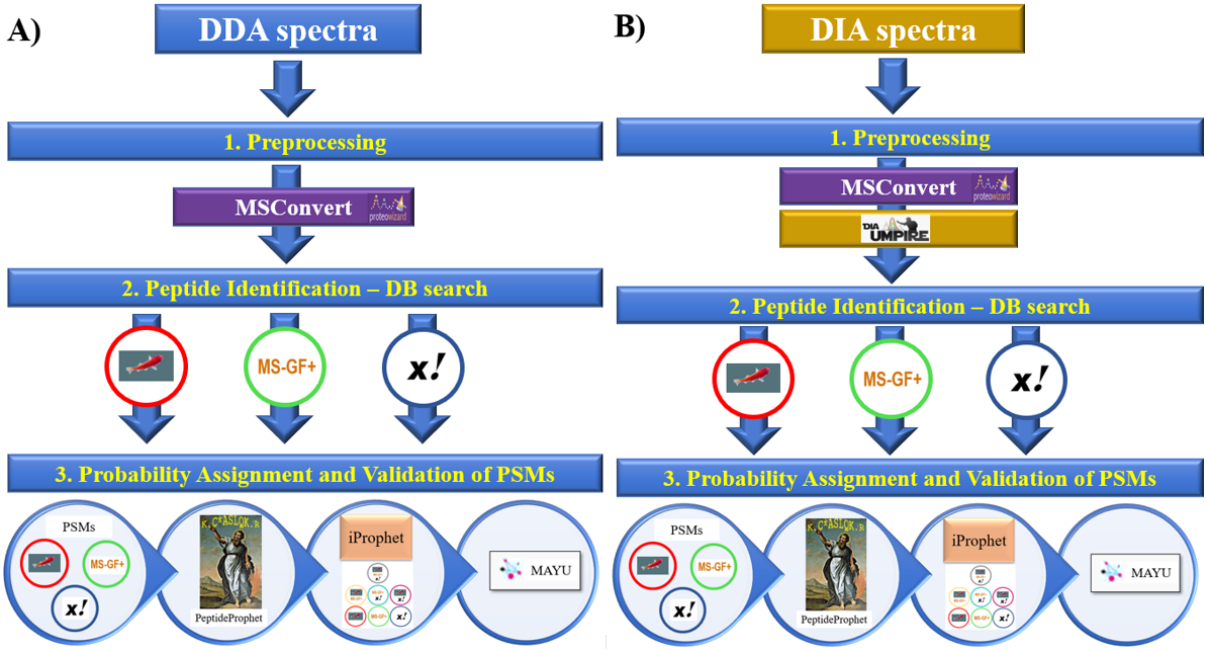


Fig. 11: Automated analysis script for DDA-based (A) and DDA-free (B) analysis.

First, MSConvert preprocesses the raw-files and gives an open-source format as output. The second step is the core of the automated script. It includes the database search of the preprocessed files with the database search engines Comet, X!Tandem, and MS-GF+. All of them differ primarily on the level of scoring function (for details see 3.5.1.3.3). It is important to note, that each database search engine runs individually and in a consecutive manner. After the individual validation step of the PSMs of Comet, X!Tandem, and MS-GF+ with PeptideProphet, the combination of individual database search engine results takes place with iProphet. To elaborate, the different individual database specific results, results of two database search engines and the PSMs of all three search tools are combined and reevaluated, respectively. Lastly, MAYU is employed for a robust FDR estimation for the corresponding results. An overview about possible combinations of search engines, abbreviations, as well as a corresponding color code is presented in Fig. 12.

Database search engine	Comet	MSGF+	X!Tandem	Comet & MSGF+	MSGF+ & X!Tandem	Comet & X!Tandem	Comet & MSGF+ & X!Tandem
Abbreviation	C	M	T	CM	MT	CT	CMT
Color code							

Fig. 12: Combinations of different database search engines for rescoring and validation with corresponding abbreviation and color code.

In total seven combinations are compared and used for further processing. In terms of flexibility the script allows individual parts of the processing pipeline to be altered, while maintaining the rest of the script. Therefore, a DDA-free, open-source DIA tool such as DIA-Umpire can be added to the script and the rest of the pipeline with all other individual programs stays constant (Fig. 11B). DIA-Umpire generates pseudo-MS/MS spectra directly from DIA data, which can subsequently be directed to database search (see details in 3.5.2.2). The database search engine step, as well as the validation steps remains the same. Hence, in theory other parts or rather other programs of the peptide identification step could be altered and the influence of the change on DIA analysis could be investigated.

6.2 Library generation and DIA analysis

Both library generation and further DIA analysis was performed with Skyline. To understand how the library size and content with identified PSMs for a given database search engine combination influences results, the in Skyline performed steps are presented in Fig. 13.

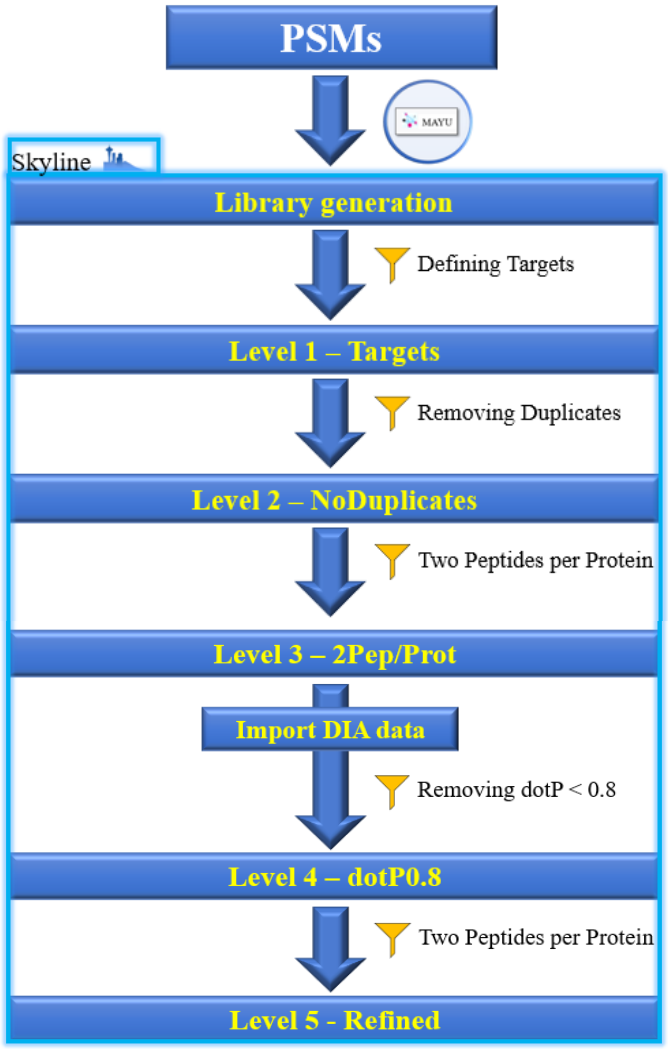


Fig. 13: Skyline analysis workflow for the database search engine combinations in detail.

For each database search engine combination, the same procedure is applied. It starts with the library generation based on the previously validated PSMs. In this context the minimum cut-off score reported by MAYU is used at which the protein FDR is under a specific threshold such as FDR < 1%. A fasta-file is imported in order to define the targets (Level 1) and duplicated peptides are removed (Level 2). Before extracting the DIA data, at least two peptides per protein are defined (Level 3). To ensure high quality data for further analysis all results with a dotP < 0.8 are removed (Level 4). For the last level only proteins with at least two peptides are included (Level 5). After exporting the results of Skyline, statistical downstream analysis is performed with MSstats and R.

7. Part II - DDA-based Analysis

7.1 Results

The seven database search engine combinations will be compared under different aspects such as library size, analysis time, file storage size, DIA data extraction, as well as the ability to identify statistically significant proteins.

7.1.1 Library size

First, the library size of the different combinations in terms of the absolute number of identified precursors (Fig. 14) and peptides (Fig. 15) is compared. The library building was performed at different error rates including an FDR of 0%, 0.5%, 1%, 1.5%, and 2%.

In Skyline only the best spectrum is chosen for a corresponding precursor from the total number of matched spectra of all DDA files. The number of precursors for the different database search engine combinations is shown in Fig. 14. On precursor-level the combination MT slightly outperforms the combination of all database search engines CMT. Both excel the other possibilities. The combinations CT, MT, and T obtain similar results, followed by M. The single variant C ranks last.

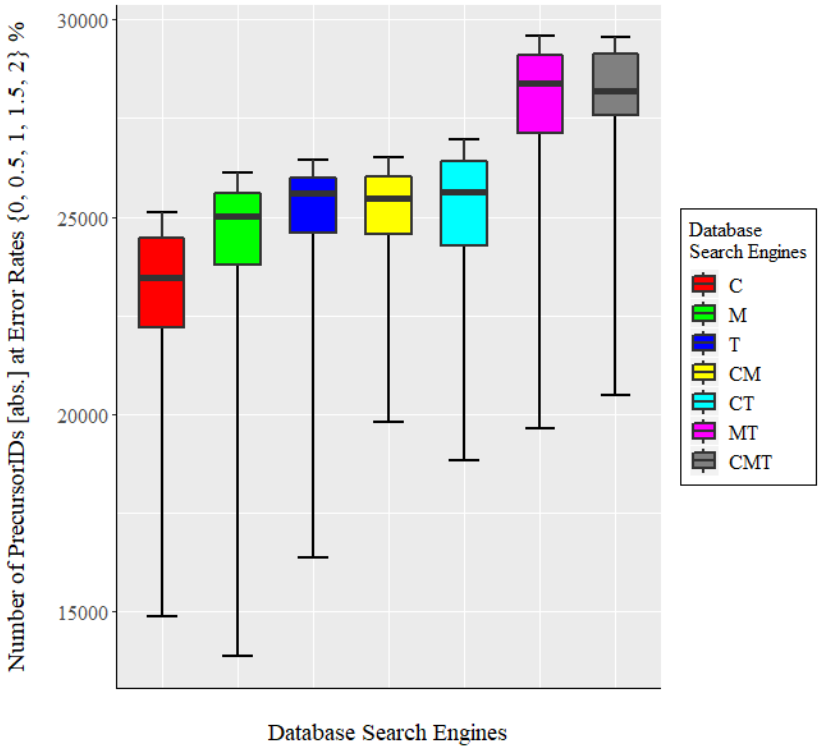


Fig. 14: Library Size - Number of precursors [abs.] at error rates {0, 0.5, 1, 1.5, 2}% for the different database search engines and combinations.

In essence, Fig. 14 indicates that starting with a specific single database search engine, the combination with results of one additional search engines always yields an increase. The step from a binary combination to a triple combination, however, is only beneficial for CM and CT and not for MT. Furthermore, the performance of a specific single database search engine can be close or even better in comparison with a binary variant, which includes two other search engines. For instance, T performs similar on an FDR of 1% than the combination CM. Hence, combining the results of multiple search engines is not always directly linked to an increase of the library size on precursor-level.

In the library precursors are assigned to specific peptides. In general, a peptide can contain a single precursor or multiple precursors. The number of peptides for the different database search engine combinations is presented in Fig. 15. The highest result on peptide-level is achieved by MT and second highest is CMT. The other possibilities are outperformed. To elaborate, the achieved number of peptides at an error rate of 0.5% by MT and CMT is higher than the obtained results at an error rate of 2% of any other option, respectively. Furthermore, the single

database search engine T and CT have similar outcomes and close behind lies CM. The single variant M has a higher number of peptides than C, which ranks last.

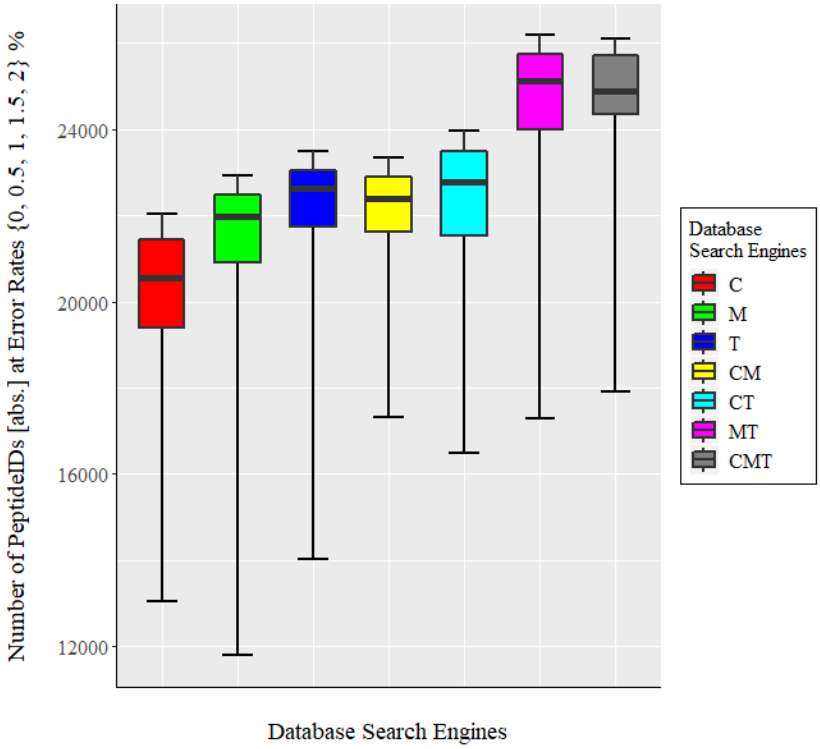


Fig. 15: Library Size - Number of peptides [abs.] at error rates {0, 0.5, 1, 1.5, 2}% for the different database search engines and combinations.

Basically, Fig. 15 shows that the behavior on peptide-level is similar to the precursor-level. The performance of a specific individual database search engine is enhanced by adding results of one further search engine. Moreover, a single variant can perform better than a binary combination as well as a binary combination can outperform the triple combination. The results indicate that combining the outcomes of multiple search engines does not necessarily result in an increase of the library size on peptide-level.

7.1.2 Analysis time and file storage size

Next, analysis time (Fig. 16) and file storage size (Fig. 17) are compared. The analysis time is the sum of the analysis time of steps two and three of the automated workflow (see Fig. 11A). The shortest time is accomplished by T with 0.53 h. In addition, C achieves a time under one hour as well with 0.86 h. The highest outcome of a single engine is obtained by M with 2.11 h,

even higher as the database search engine combination CT with 1.92 h. The combination of all three database search engines needs 3.44 h, followed by MT with 3.23 h and CM with 2.98 h.

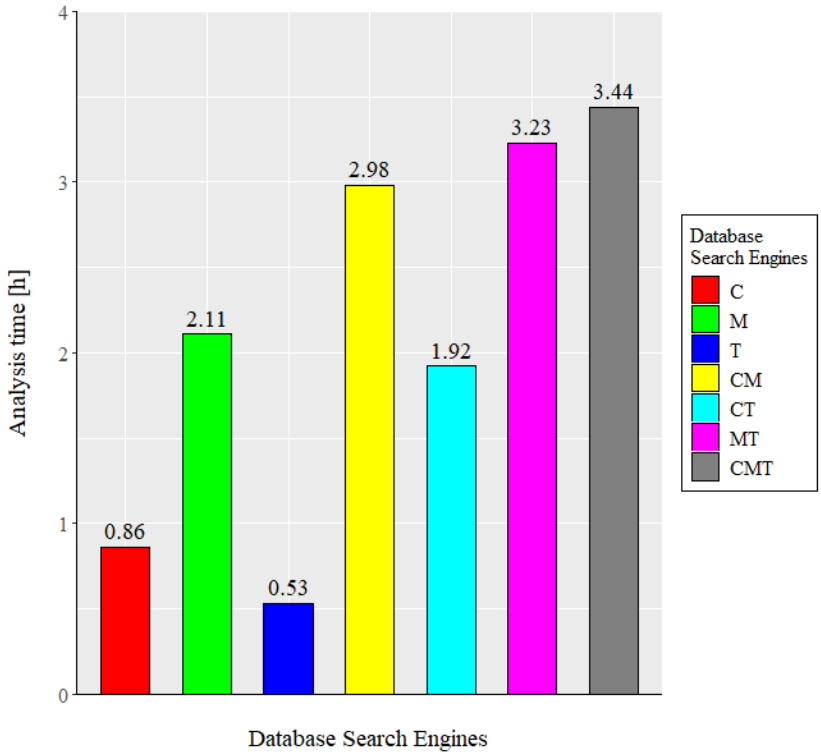


Fig. 16: Analysis time [h] for the different database search engine combinations.

On the one hand, Fig. 16 shows that adding a search engine result to a specific search engine always increases the analysis time. For example, if C is combined with T the analysis time goes up from 0.86 h to 1.92 h. And if C is combined with T and M, it increases to 3.44 h. On the other hand, a generalization that the combination of multiple search engines will always directly lead to an increased analysis time in comparison with any single search engines is not possible. The analysis of M and corresponding combinations takes the longest. In comparison both C and T need relatively short times. As a result, the combination CT has a smaller analysis time than the single search tool M.

The file storage size covers the sum of every file generated starting from the peptide identification step and ends after the statistical validation step with MAYU (see Fig. 11A). In detail, CMT requires 3.97 GB storage size. The binary combinations CM and CT obtain results of 3.18 GB and 3.16 GB, respectively. Next in the order is the single database search engine C

with 2.37 GB. The combination MT needs a file storage size of 1.94 GB and the smallest requirements include M with 1.11 GB and T with 1.10 GB.

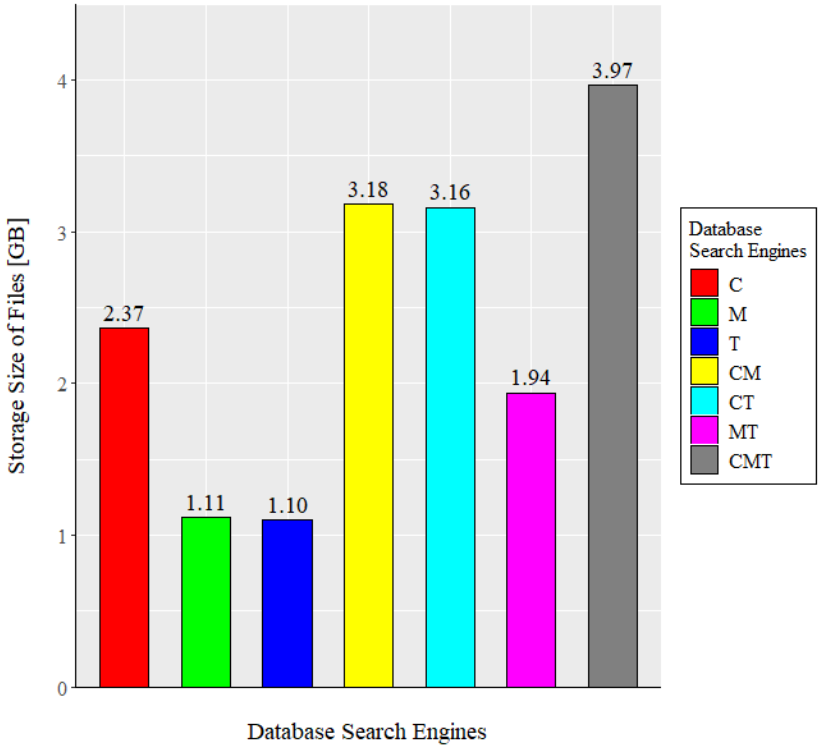


Fig. 17: Storage size of files [GB] for the different database search engine combinations.

The results in Fig. 17 indicate, that adding results of search engines to a specific single search tool always yields an enhanced storage size. In contrast, an excellent performance of single search engines and their combination can result into the fact that a single search tool requires more space than a binary combination. For example, the combination MT outperforms C.

7.1.3 Data Mining – downstream analysis & SWATH quantification performance

In the following chapter, the downstream analysis in Skyline (see Fig. 13) and its effect on protein- and peptide-level for the corresponding libraries generated with an FDR < 1% will be examined. In brief, Level 1 refers to the target definition, Level 2 is based on removing duplicates, Level 3 restricts further analysis to two peptides per protein prior to DIA-data import, Level 4 removes peptides with a dotP < 0.8, and lastly on Level 5 again a restriction for two peptides per protein is performed. The development will be displayed in absolute numbers across the downstream analysis. In this context, the SWATH quantification performance is of

special interest, which refers to high quality assignments based on a dotP < 0.8 after the DIA data extraction (transition from Level 3 to Level 4). For further illustration of the impact of individual filter steps and to investigate the benefit of combining multiple search engines, the development of the ranking order based on the performance of individual database search engine combinations will be presented. In addition, the similarity of identifications by different search engines will be investigated.

7.1.3.1 Downstream analysis on protein-Level

The development of the absolute number of proteins for the respective database search engine combination is presented in Fig. 18. Starting in the range between 5000 and 5500 protein identifications on Level 1, the number drops about 20% on Level 2 and 40% to approximately 3000 proteins on Level 3. The biggest loss of proteins happens from Level 3 to Level 4 to around 200 proteins per database search engine. This corresponds to a decline of nearly 95% relative to Level 1. The last filtering step leads to around 100 proteins. Hence, the total number of detected proteins descends around 98% from Level 1 to Level 5 for every search tool.

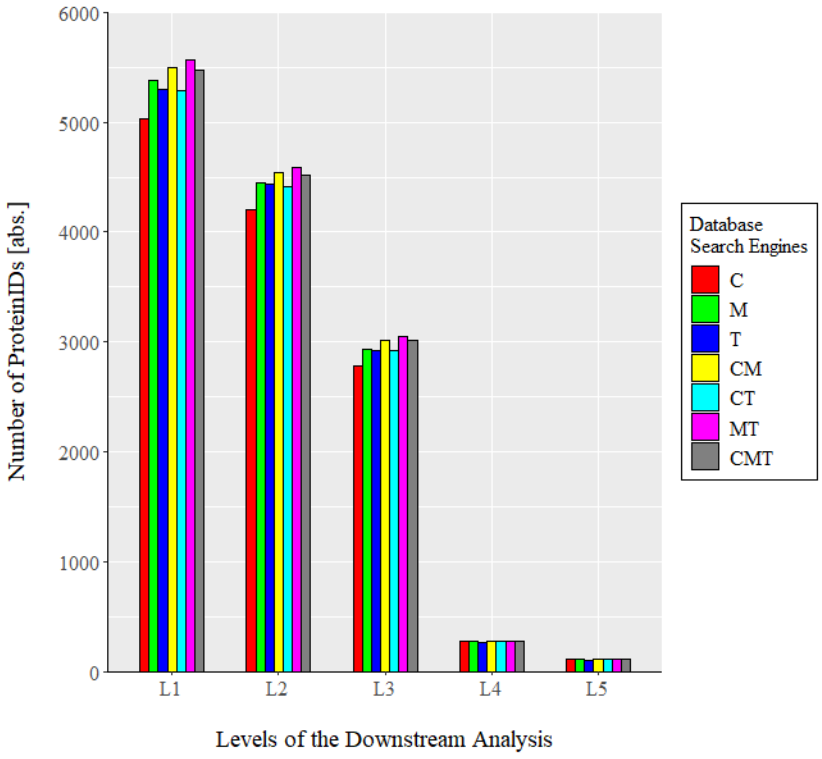


Fig. 18: Development of the number of ProteinIDs [abs.] during downstream analysis for the different database search engine combinations.

To further evaluate the consistency of the filtering steps, the ranking based on protein identifications is depicted in Fig. 19. The best performing option has the highest number of identifications and ranks first for a given level. If database search engines achieve the same ranking for an analysis level, the following rank is omitted. However, the ranking remains constant from Level 1 to Level 2 with MT ranking first and C ranking last. While the change to the next level introduces small changes, the filtering from Level 3 to Level 4 affects the ranking drastically. Both MT and CMT drop down to rank 5 and C attains rank 3. The next step to Level 5 goes along with small changes resulting into a leading performance of C and CT, which rank place 7 and 6 at Level 1, respectively.

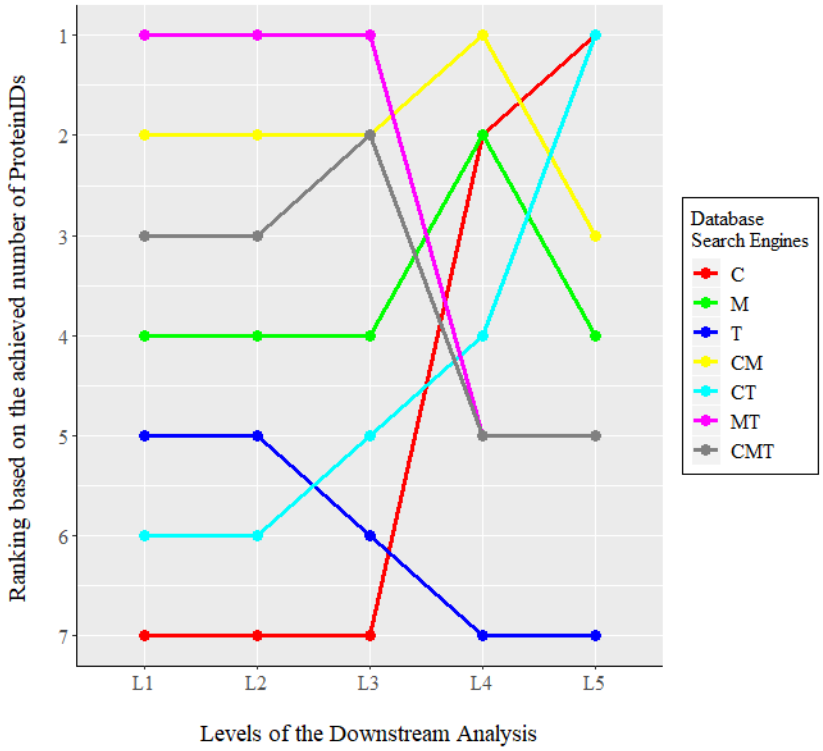


Fig. 19: Ranking based on the achieved number of ProteinIDs for the different database search engine combinations during downstream analysis. Libraries generated with an FDR < 1%.

7.1.3.2 Downstream analysis on peptide-level

The influence of the downstream analysis on peptide-level in terms of absolute numbers is shown in Fig. 20. The initial number of peptides drops from around 24000 to 400 identifications. This correlates with a decrease of 98% from Level 1 to Level 5. Especially the

transition from Level 3 to Level 4 contributes to the drastic reduction in identifications for each database search engine.

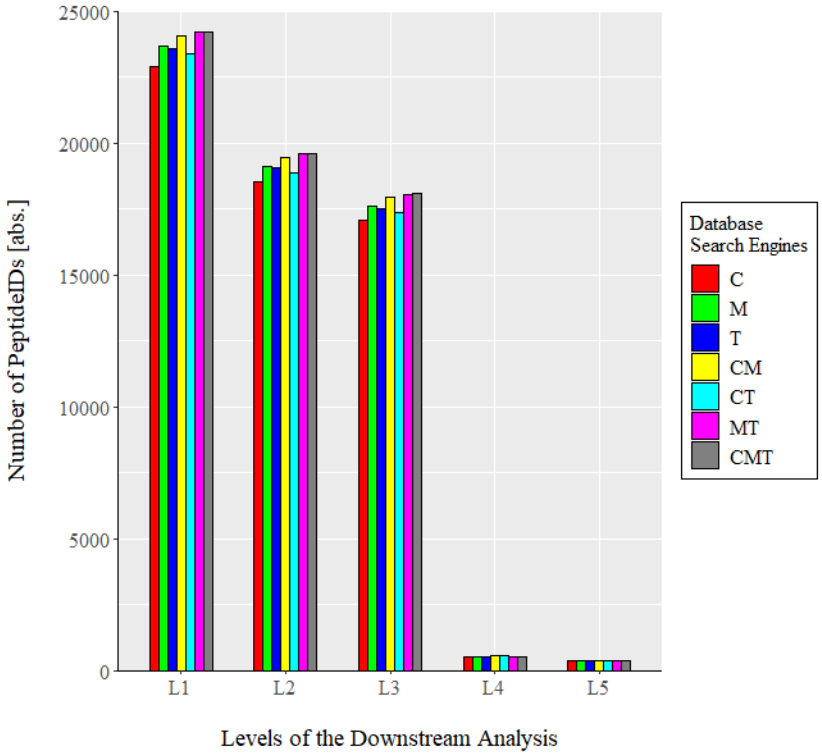


Fig. 20: Development of the number of PeptideIDs [abs.] during downstream analysis for the different database search engine combinations.

The ranking based on protein identifications is shown in Fig. 21. While the order stays mainly constant from Level 1 to Level 3 with CMT and MT at the top and CT and C at the bottom, the transition from Level 3 to Level 4 changes the ranking significantly. The options C and CT improve their performance and CMT and MT decline to rank 5 and 6, respectively.

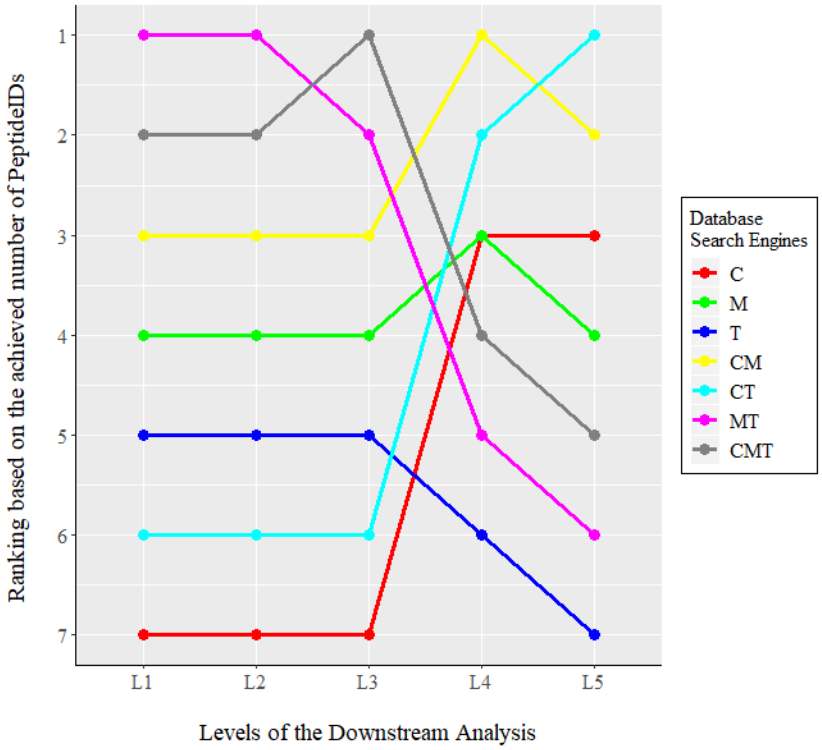


Fig. 21: Ranking based on the achieved number of PeptideIDs for the different database search engine combinations during downstream analysis. Libraries generated with an FDR < 1%.

7.1.3.3 Analysis of the influence of signal intensity and retention time variation on SWATH quantification performance

The previous reported results demonstrate a drastic decline of identifications from Level 3 to Level 4. To evaluate the DIA extraction further, the assigned signal intensity of the transitions stored in the library, as well as the coefficient of variation (CV) of the retention time of transitions are compared between Level 3 and Level 4. In brief, the DIA data is imported on Level 3 and then low-quality data (dotP < 0.8) is excluded leading to Level 4. First, the signal intensities of transitions in the library are extracted for the assignments, which are present at Level 3. Second, the same procedure is applied for Level 4. Next, the extracted library signal intensities are averaged per precursor, respectively.

The corresponding averaged transition signal intensities extracted from the library, which have a protein assignment on Level 3 are displayed in Fig. 22 and consecutively the averaged transition signal intensities extracted from the library, which have a protein assignment on Level 4 are shown in Fig. 23. Both comparisons are presented via boxplots. Note, that outliers are not displayed to achieve a better overview.

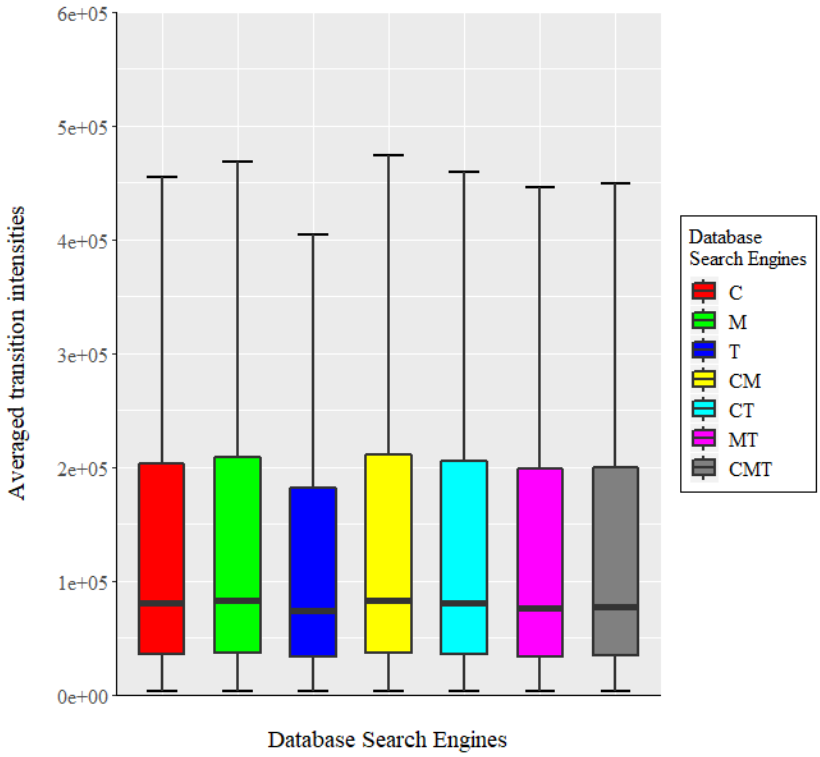


Fig. 22: Averaged transition intensities for different database search engine combinations, which are stored in the library and assigned to proteins on data analysis Level 3. Outliers are not shown.

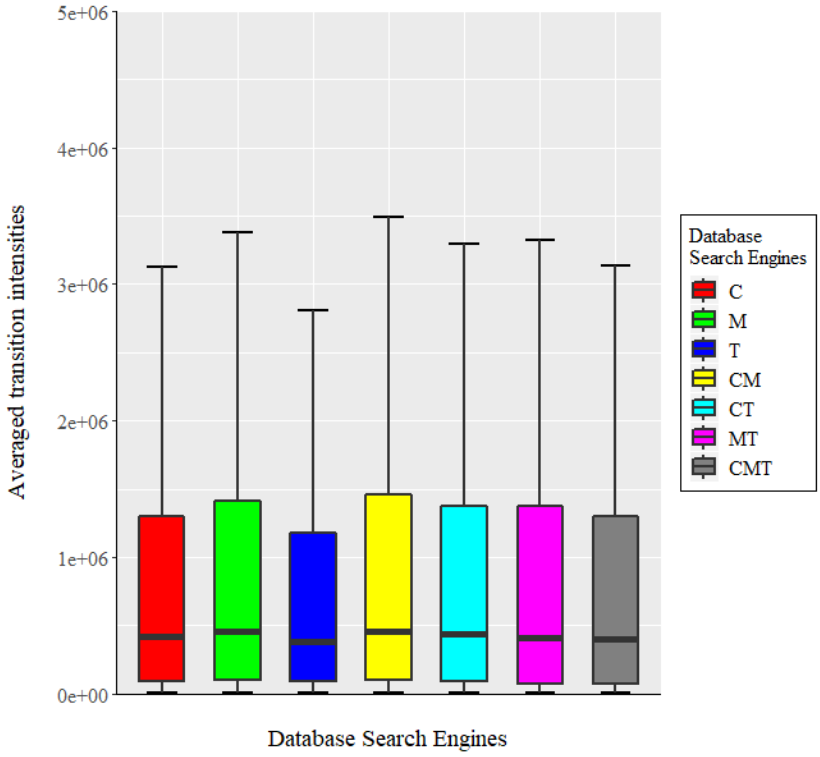


Fig. 23: Averaged transition intensities for different database search engine combinations, which are stored in the library and assigned to proteins on data analysis Level 4. Outliers are not shown.

In Fig. 22 the transitions, which are stored in the library and assigned on Level 3, display a median signal intensity from the lowest value 3.79e+05 for T to the highest median 4.53e+05 of M and CT. In contrast, Fig. 23 shows that on Level 4 the median ranges between 7.40e+04 for T to a value of 8.26e+04 for M. This comparison suggests that especially precursors with low signal intensities correspond to a dotP < 0.8 and are removed from Level 3 to Level 4.

Furthermore, the CV of retention times for each transition is determined and averaged per precursor for each database search engine combination. The results for Level 3 are displayed in Fig. 24 and for Level 4 in Fig. 25.

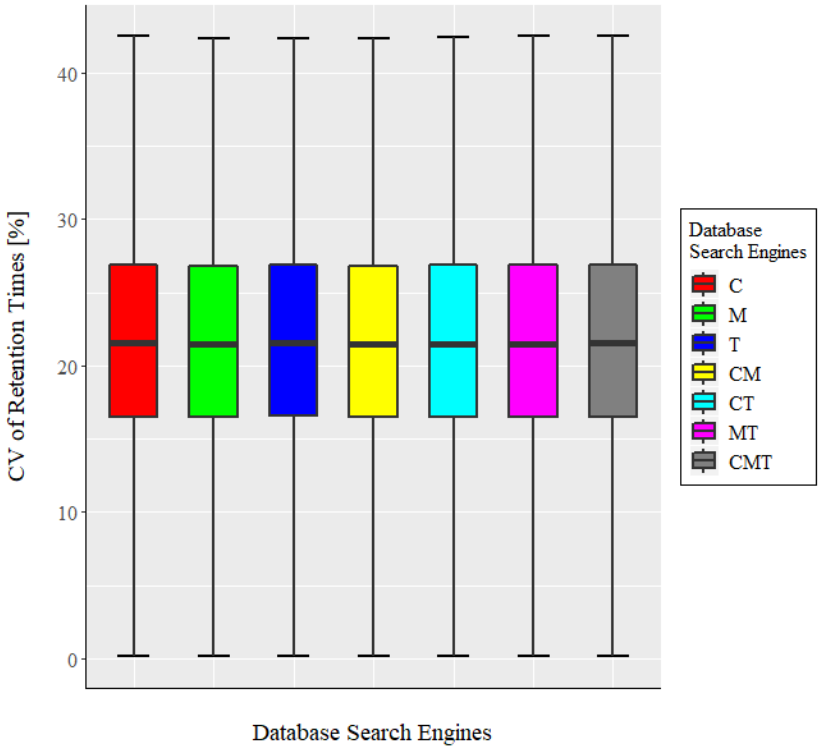


Fig. 24: CV of retention times [%] per precursor for different database search engine combinations on Level 3. Outliers are not shown.

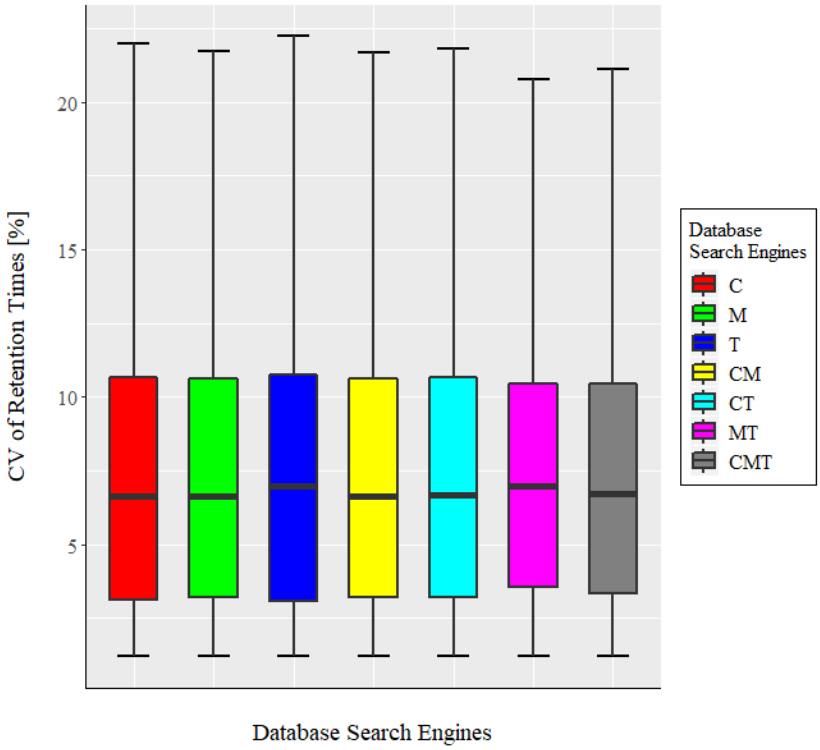


Fig. 25: CV of retention times [%] per precursor for different database search engine combinations on Level 4. Outliers are not shown.

While the median of the CV of the retention times on Level 3 is approximately 21% across every database search engine (Fig. 24), the median on Level 4 is leveling off at 7% for each search tool option (Fig. 25). The comparison between Level 3 and Level 4 indicates that low CVs of retention times correlate with high-quality DIA data extraction.

To sum up, the findings insinuate that both low-abundant transitions and a high retention time variability are prone to low-quality DIA data extraction.

7.1.4 Analysis of the consistency of detecting statistically significant proteins

This chapter focuses on the consistency of detecting statistically significant proteins across the database search engine combinations. Biological inference with concentration on the identity of individual proteins and possible biological patterns is performed in "Part V – Biological Inference". However, to evaluate the consistency the total number of statistically significant proteins in stage-wise comparisons is determined and the similarity between the findings is studied. A detailed view via volcano plots for each stage-wise comparison for the corresponding

database search engine combinations is presented in the appendix (Fig. 56 - Fig. 79). An FDR of max. 5% is applied as cut-off to determine statistical significance.

Furthermore, it is important to note that after the data analysis workflow in Skyline ("Level 5 – Refined"; Fig. 13) the number of proteins, which are subjected to statistical analysis, differs among the database search engine combinations. An overview of the number of proteins for the corresponding database search engines on Level 5 is presented in Fig. 26. The single variant C obtains 116 proteins, as well as the combination CT. Next in the ranking is CM with 114 proteins, followed by M with 111. Both the combination of all three database search engines CMT and MT achieve 109 proteins. The single database search engine T has 108 proteins which are submitted for further statistical analysis.

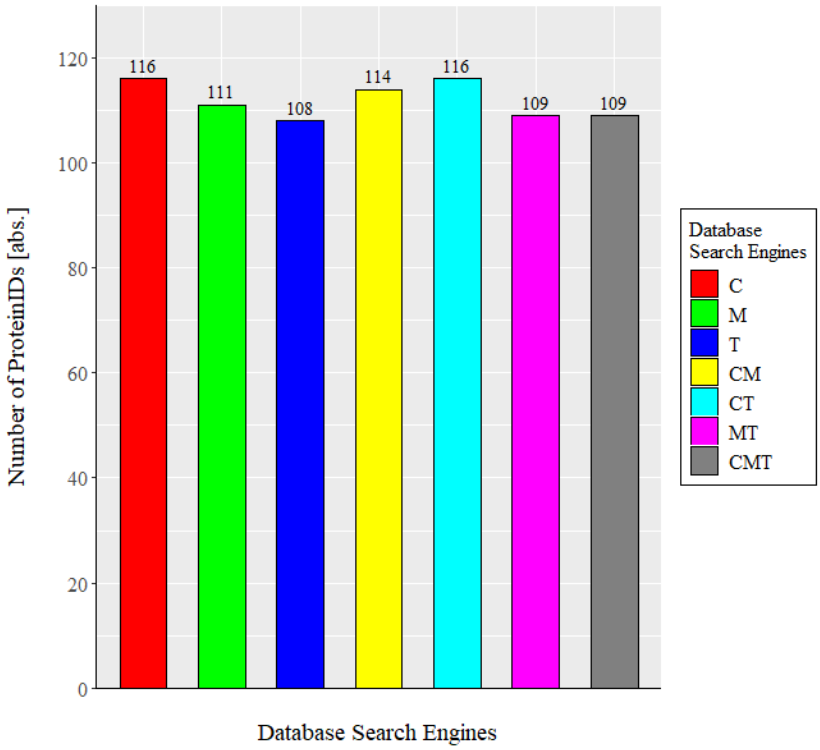


Fig. 26: Number of proteins [abs.] after Skyline analysis for the different database search engine combinations.

To examine the similarity of results in detail, the coverage of proteins after Skyline analysis is presented in Fig. 27. Every protein of each database search engine combination is combined, the duplicated proteins are removed, and the total number of unique elements is determined (122 proteins). The coverage of proteins of a database search engine displays the proportion of

detected proteins in comparison with the number of total unique elements. It is important to note that the same percentage of coverage does not necessarily imply that the same proteins are present; it only indicates that the absolute number of proteins is the same.

Analogous to the absolute number of proteins C and CT perform best and achieve a coverage of 95%. Next in the ranking is M the 91%, closely followed by T, MT and CMT with 89%. The results of Fig. 27 show that mainly the same proteins are present for statistical analysis of the different database search engines. Furthermore, none of the possibilities achieves 100%.

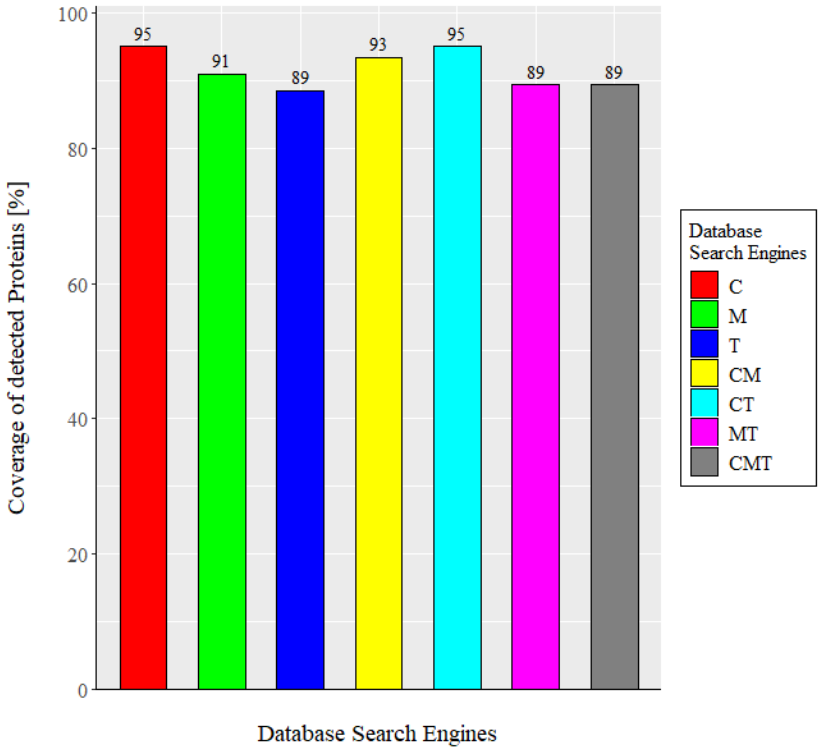


Fig. 27: Coverage of proteins [%] after Skyline analysis for the different database search engine combinations.

Next, for each stage-wise comparison and respective database search engine the absolute number of statistically significant proteins and the corresponding coverage is shown in table 1. Note, that statistically significant findings are only detected for the stage-wise comparisons SI vs. SIV, SII vs. SIII, and SII vs. SIV. Additionally, the total number of statistically significant proteins for the different database search engine combinations are summed up for all stage-wise comparisons (depicted as "Total" in table 1). In detail, the findings of each stage-wise comparison for a corresponding database search engine are combined and duplicated

proteins are removed. To elaborate, if for a specific database search engine a protein is statistically significant in the comparison between SI vs. SIV and for example in the comparison between SII vs. SIII, it will be counted as one. For the corresponding coverage all detected statistically significant proteins are combined and duplicates removed resulting in 22 unique proteins across each stage-wise comparison and each database search engine.

Table 1: Number of statistically significant hits (FDR < 5%) and coverage for stage-wise comparisons and the respective database search engine.

Database search engines	**SI vs. SIV** ProteinIDs [abs.]/ Coverage [%]	**SII vs. SIII** ProteinIDs [abs.]/ Coverage [%]	**SII vs. SIV** ProteinIDs [abs.]/ Coverage [%]	**Total** [a)] ProteinIDs [abs.]/ Coverage [%]
C	7/70	11/79	3/75	18/82
M	8/80	8/57	3/75	15/68
T	6/60	12/86	3/75	18/82
CM	8/80	11/79	3/75	18/82
CT	10/100	10/71	4/100	18/82
MT	7/70	9/64	3/75	16/73
CMT	10/100	9/64	4/100	17/77

[a)] "Total" refers to the combination of statistically significant protein of each stage-wise comparison for a respective database search engine excluding duplicate proteins.

Table 1 shows that in the stage-wise comparison SI vs. SIV and SII vs. SIV the database search engines CT and CMT achieve a coverage of 100% with detecting 10 and 4 statistically significant proteins for the respective comparison. For SII vs. SIII no database search engine combination is able to identify all statistically significant hits. The closest search tool is T with 12 significant findings corresponding to a coverage of 86%. In addition, no database search engine detects the total number of statistically significant proteins. The search engines C, T, CM, and CT all obtain 18 proteins, which refers to a coverage of 82%. Hence, no search engine is able to cover all statistically significant hits.

In total, the results of the stage-wise comparisons demonstrate an inherently consistency of detecting statistically significant proteins. However, the question about why a certain database search engine combination performs better than the other and if there is a possible connectivity between the library size and statistically significant hits will be addressed in the next chapter.

7.2 Discussion

In the following chapter, all previously presented results regarding library size, analysis time, data storage size, downstream analysis, and statistically significant hits will be evaluated and discussed in a more detailed fashion. Moreover, individual categories will be put into context to one another to highlight possible dependencies and examine potential benefits of combining the results of multiple database search engines.

7.2.1 Library size, analysis time, and storage size

Regarding the library size the performance on peptide- and precursor-level of a specific individual database search engine is enhanced by adding results of one further search engine. Moreover, a single variant can perform better than a binary combination. In addition, a binary combination can outperform the triple combination. In general, the performance on peptide- and precursor-level depend among other factors mainly on the performance as single database search tool and the complementarity between database search engines within combinations. Complementarity refers to the fact that the overlap between sets of spectra of different database search engines is significantly less, if the search tools employ different scoring strategies. Combining database search engines relying on distinct scoring functions is most beneficial to increase the identification rate (see 3.5.1.3.3) [66]. Both the search engines C and T apply a descriptive model to access peptide spectrum matches. The search tool M is based on an interpretative strategy. In theory, CM and MT should outperform CT. Nevertheless, because T performs significantly better than C and M as an individual search tool, the combination between CT and T on its own is better than the combination CM. However, the binary combination MT performs best on precursor- and peptide-level regarding the library size. Furthermore, it is observed that adding the results of one database search engine to another search tool always increases the identification rate on library-level. In addition, it is important to mention that it is has been reported that different database search tools perform best on distinct datasets and conditions [66]. Thus, the findings do not serve to generalize the individual performance of the used database search engines.

In terms of analysis time and storage size, the single variant T performs best. Between the binary combinations, CM ranks last. If the analysis time is valued more, CT achieves better results

than MT. If the focus is on storage size, MT is favored. It is clear, that an enhanced number of utilized search tools goes along with an increased analysis time and storage space. Nevertheless, with the development of cloud based computing and the opportunity to store files, as well as to process several search files at a time, additional computational resources will become less important [66]. However, a favorable tradeoff between identification rate, storage size and analysis time is accomplished by the single search engine T. In regard to identifications T gets to the range of the binary combinations CM and MT. Moreover, in terms of analysis time and storage size T performs best out of all combinations. With the aim in mind to enhance the identification rate on library-level, the combination MT ranks first. The increase on peptide- and precursor-level regarding identifications trumps the enhanced analysis time.

7.2.2 SWATH quantification performance and reproducibility of the detection of statistically significant proteins

The impact of the downstream analysis on protein- and peptide-level for each database search engine was investigated. After performing all filtering steps in the downstream analysis up to 98% of the determined identifications are excluded, respectively. The biggest decrease is based on a poor SWATH quantification performance. In the transition from Level 3 to Level 4, in which all proteins that obtain a dotP < 0.8 are removed, only 2% remain and display high-quality assignments. Additional data mining insinuated that both low-abundant transitions and a high retention time variability are prone to low-quality DIA data extraction. Moreover, this transition alters the performance-based ranking of the different database search engine combinations. In detail, while MT achieves the highest results on library-level and after target import (Level 1), the performance declines to rank 5 on Level 5. Hence, the gained information input on library-level of MT is not only drastically reduced, but also results into a lower performance in comparison with other database search engine combinations. As a conclusion, the results indicate that for DIA data extraction the quality of the information input of the library is more important than the mere total number of identifications stored in the library. To sum up, while combining multiple database search engines enhances the sensitivity on library-level it does not guarantee an adequate DIA data extraction. By comparing the different database search engines, it is obvious that the increase on library-level is not proportional to the absolute number of proteins and peptides extracted from the DIA data. However, to connect a certain library spectrum of a specific database search engine to an achieved dotP for a corresponding

protein and examine its influence on the downstream analysis in the context of the whole composition of the library input is beyond the scope of this project. Again, it is noteworthy, that all made considerations are primarily limited to this dataset.

The results of the statistical analysis show an inherent consistency for all database search engines. Best performing combination regarding the coverage of the most statistically significant hits is CT. The most promising combination on library-level MT is ranged at the bottom of the ranking. It is clear, that subtle changes on product-, precursor-, peptide-, and protein-level lead to distinct results. As a result, the choice of library directly impacts the detection of statistically significant proteins. In addition, the results of the statistical analysis demonstrate that no database search engine obtains a total coverage of 100% of statistically significant findings. In other terms, performing an analysis only based on a single database search engine and the resulting library will not exploit the total amount of information given by DIA data. In addition, if several database search engines and the resulting library lead to detecting the same statistically significant protein, it adds confidence and verifies the findings.

7.3 Conclusion

The DDA-based data analysis workflow aimed at boosting the identification rate on library-level and subsequently the sensitivity of the DIA analysis by combining the results of multiple database search engines for library generation. An overview of the benefits and drawbacks of every database search engine combination is provided in table 2.

In all cases the identification rate on library-level was increased by combining results of one search engine to another to form a binary combination. Merging the results of search engines was correlated with an increased analysis time and storage size. In terms of the library size, the combination MT excelled the other possibilities. Remarkably, the increase of the library input was not proportional to the extracted protein and peptides. Regarding the number of extracted high-quality assignments the option CT ranked first. As a result, an increased identification rate on library-level did not guarantee an enhanced sensitivity of the DIA analysis. Further investigation directed the attention on the retention time variability of the transitions and the quality of library input regarding signal intensity to improve the DIA data extraction. Moreover, statistical evaluation has demonstrated that no option covers the total amount of statistically

significant proteins. The database search engine combination CT obtained the most statistically significant results based on an FDR threshold of max. 5%.

Table 2: Overview of merits and drawbacks of each database search engine combination.

Categories	C	M	T	CM	CT	MT	CMT
Analysis time	++	+	++	-	+	-	-
Storage size	+	++	++	-	-	+	-
Number of precursors in Library (FDR < 1%)	-	+	+	+	+	++	++
Number of peptides in Library (FDR < 1%)	-	+	+	+	+	++	++
Number of extracted proteins (FDR < 1%) [a)]	++	+	-	+	++	-	-
Extraction of statistically significant proteins	+	-	+	+	+	-	+

[a)] Refers to Level 5 – after all filtering steps are performed (see Fig. 13).

As a conclusion, it is recommendable to use the results of two database search engines and run the total analysis both for the search tools individually and their binary combination. In this manner, the chances to extract valuable information provided by the DIA data is increased, it adds certainty to the findings, and thus enhances the probability to achieve the main goal of the thesis – elucidating biological significant proteins and pathogenic patterns for CRC in the sense of systems biology.

8. Part III - DDA-free Analysis

8.1 Results

After applying the data analysis based on the DDA-free approach, the seven database search engine combinations will be compared in different aspects – library size, analysis time, file storage size, consistency of the data analysis, DIA data extraction and reproducibility in identifying statistically significant proteins.

8.1.1 Library size

Initially, the different database search engine options are compared regarding the absolute number of precursors (Fig. 28) and peptides (Fig. 29) stored in the library. The library generation was carried out at various error rates including an FDR of 0%, 0.5%, 1%, 1.5%, and 2%.

The number of precursors for the different database search engine combinations is presented in Fig. 28. At an FDR of 1% the single variant T performs best, closely followed by the binary option CT and the combination CMT. Next in the ranking is MT. While M performs worse, C and CM obtain similar results.

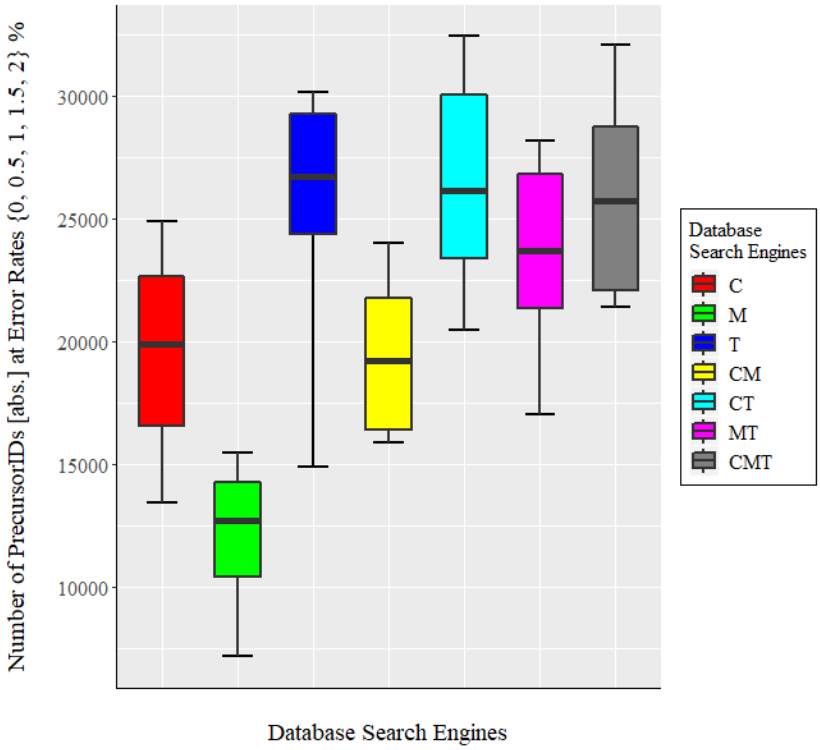


Fig. 28: Library Size - Number of precursors [abs.] at error rates {0, 0.5, 1, 1.5, 2}% for the different database search engines and combinations.

The results presented in Fig. 28 show a significant difference between the single search engines. Especially, T excels and M performs significantly worse than the other possibilities. Adding results to C or M to form a binary combination always yields an enhancement of the library size on precursor-level. Further combination to a triple combination does not assure a growth of library input. It is only advantageous for CM and MT but not for CT. As a result, a single variant can outperform a binary combination and a binary combination can perform better than the triple combination. Hence, joining the results of multiple search engines is not always connected to an increase of the library size.

The number of peptides for the different database search engine combinations is presented in Fig. 29. The results on peptide-level are in accordance with the outcomes on precursor-level. Again, T excels every other option at an FDR of 1%. The combinations MT and CMT are close behind. The binary combination MT is next in the ranking. Moreover, the single database search engine C and CM have similar outcomes and M performs worse.

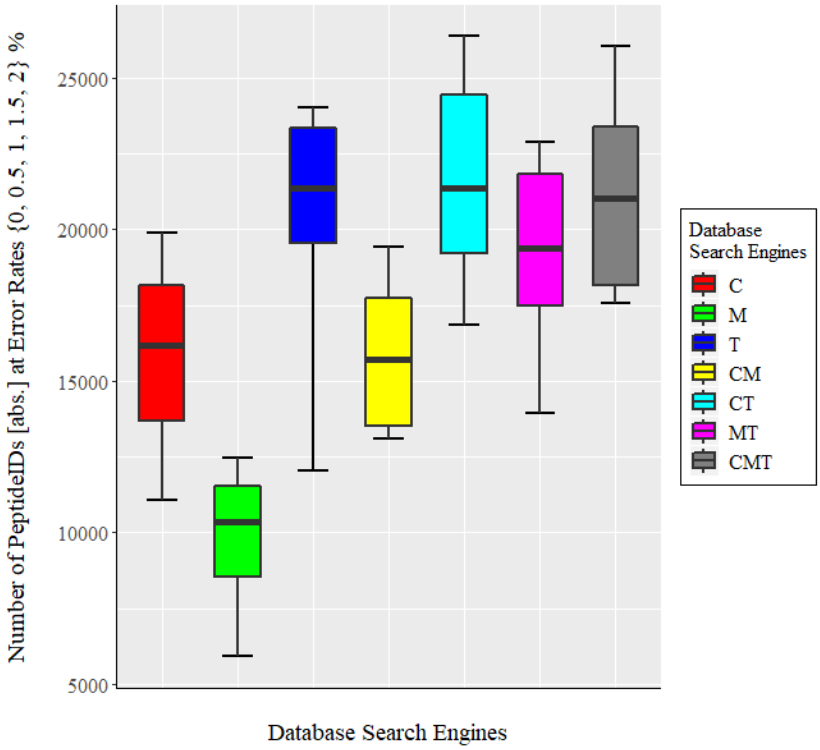


Fig. 29: Library Size - Number of peptides [abs.] at error rates {0, 0.5, 1, 1.5, 2}% for the different database search engines and combinations.

In essence, Fig. 29 indicates that the performance of a specific individual database search engine is enhanced by adding results of one further search engine. Only the single variant T outperforms every other option contributing to the fact that a single search engine can achieve a higher library input than a binary combination. The findings show that combining the outcomes of multiple search engines is not necessarily beneficial regarding an increase of the library size.

8.1.2 Analysis time and file storage size

In the following chapter, analysis time (Fig. 30) and file storage size (Fig. 31) are analyzed. The analysis time is the sum of the analysis time of step two and three of the automated workflow including the DIA-Umpire module (see Fig. 11B). Hence, in comparison to the DDA-based analysis, now the generation of pseudo-MS/MS spectra contributes to the analysis time.

The data analysis of the single engines M and T both need approximately 63.0 h, performing best in comparison with all other possibilities. Next, in the ranking is C with 75.4 h. The binary

combination MT obtains a time of 96.8 h, CM lasts 105 h and CT needs 113 h. The triple combination CMT takes the longest time with 145 h.

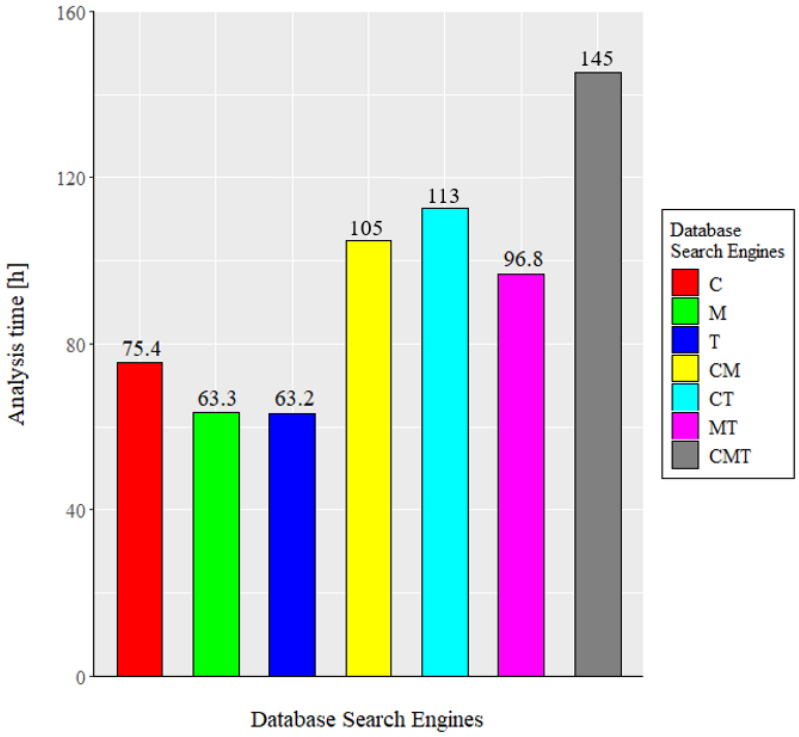


Fig. 30: Analysis time [h] for the different database search engine combinations.

The analysis times in Fig. 30 display that combining the results of multiple search engines always leads to an increase in analysis time. From a single search engine, the analysis time increases up to 65% to a binary option and around 100% to the triple combination. In addition, the step from a binary combination to CMT leads to a growth around 45%. Hence, merging results of multiple search engines has a significant effect on the analysis time.

The file storage size sums up each generated file from step two and three of the automated workflow including the generation of pseudo-MS/MS spectra by DIA-Umpire (see Fig. 11B). The triple combination CMT needs 95.4 GB storage size. The binary options CM and CT require 75.6 GB and 85.2 GB, respectively. Moreover, the single variant C has 64.4 GB, closely followed by MT with 62.8 GB. Next in the ranking is T with 52.5 GB and the smallest requirement is achieved by M with 42.4 GB.

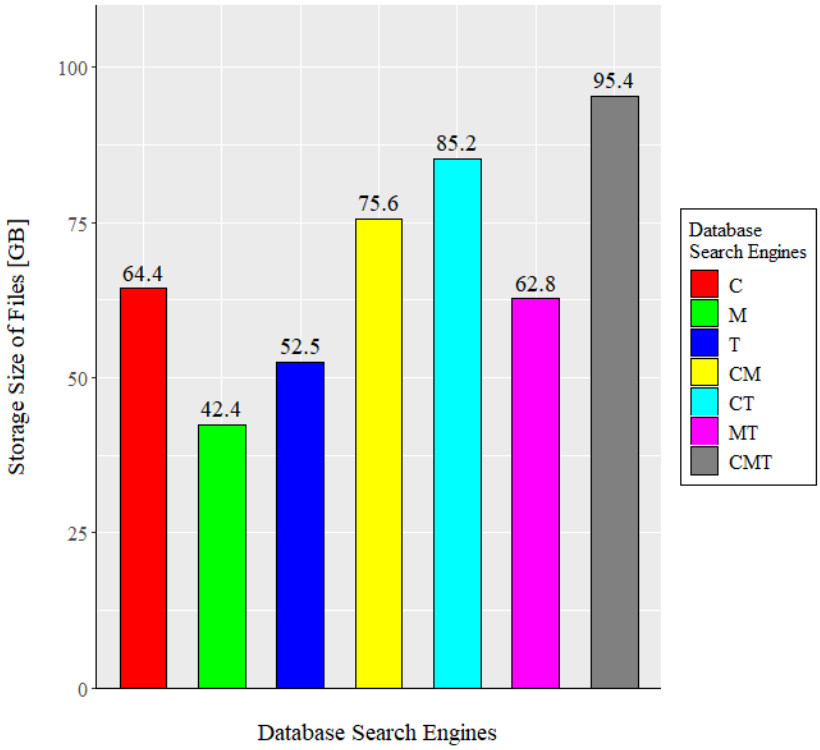


Fig. 31: Storage size of files [GB] for the different database search engine combinations.

From Fig. 31 it is obvious that adding results of search engines to a specific search tool always leads to an increased memory requirement. On the other hand, an outstanding achievement of two individual search tools can lead to the outcome that their binary option performs better than a different single search tool. For example, the combination MT requires less storage size than the single search engine C.

8.1.3 Data Mining – downstream analysis & SWATH quantification performance

In this chapter, the data analysis in Skyline (see Fig. 13) and the influence on protein- and peptide-level for the respective libraries (FDR < 1%) will be investigated. In brief, Level 1 includes the target definition, for Level 2 duplicated peptides are removed, Level 3 limits further analysis to two peptides per protein prior to DIA-data import, Level 4 filters peptides with a dotP < 0.8, and on Level 5 again a restriction for two peptides per protein is carried out. The development will be examined in absolute numbers of the identifications. Furthermore, the SWATH quantification performance is of particular interest, which refers to DIA data extraction and quality refinement based on a dotP < 0.8. In order to illustrate the impact of individual filter steps and to highlight potential advantages of joining multiple database search

engines, the development of the ranking of individual database search engine combinations in terms of the achieved number of absolute identifications will be displayed. Moreover, the similarity of the findings by different search engines will be studied and the consistency of the data analysis will be evaluated.

8.1.3.1 Downstream analysis on protein-level

The dependency of the downstream analysis on the absolute number of proteins for the respective database search engine combination is shown in Fig. 32.

Every search engine option achieves around 3000 protein identifications on Level 1, except M with approximately 2000. During the data analysis the results decline about 90% from Level 1 to Level 5 resulting into approximately 300 proteins for each database search engine combination. The highest decrease occurs in the transition from Level 3 to Level 4.

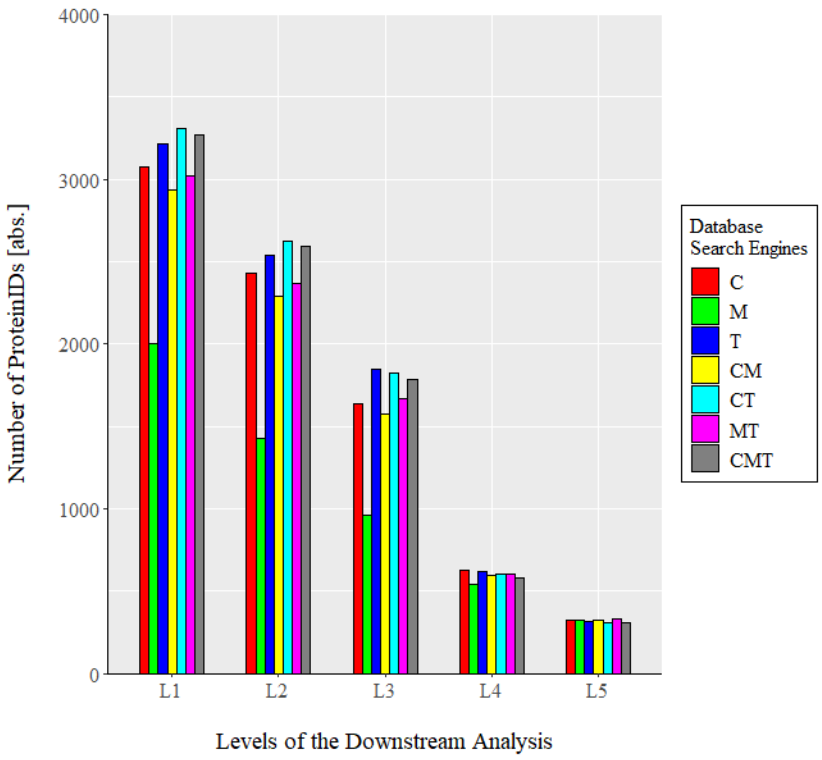


Fig. 32: Development of the number of ProteinIDs [abs.] during downstream analysis for the different database search engine combinations.

Moreover, the influence of the downstream analysis on the number of identifications per database search combination is studied (Fig. 33). A ranking is determined for each analysis level, in which the best performing possibility has the highest number of identifications and ranks first for an individual level. If database search engines obtain the same ranking for an analysis level, the following rank is omitted. However, the first transition from Level 1 to Level 2 has no impact on the ranking, in which CT ranks first and M ranks last. While the following filter step from Level 2 to Level 3 results into small alterations, the next steps introduce significant changes. For example, CT declines to rank 6 at Level 5 and MT ranks first starting at rank 5 at Level 1.

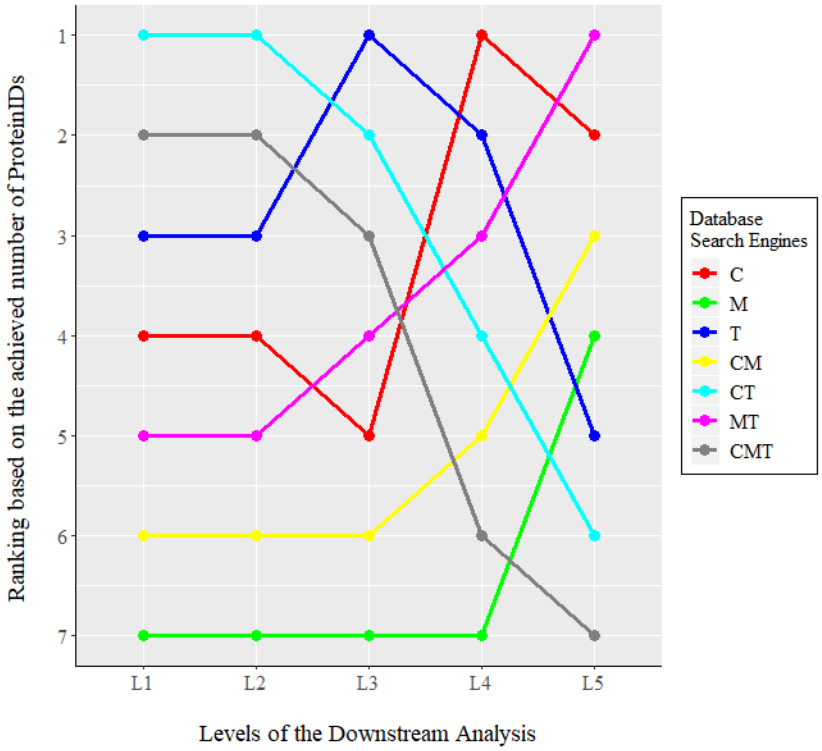


Fig. 33: Ranking based on the achieved number of ProteinIDs for the different database search engine combinations during downstream analysis. Libraries generated with an FDR < 1%.

8.1.3.2 Downstream analysis on peptide-level

Next, the impact of the downstream analysis on peptide-level is examined. The development of peptide identifications is presented in Fig. 34. In accordance with previous findings on protein-level, peptide identifications decline to 90% beginning with around 20000 identifications on

Level 1 and resulting to nearly 1500 results at Level 5 for the different database search engine combinations. Again, the highest decrease corresponds to the transition from Level 3 to Level 4.

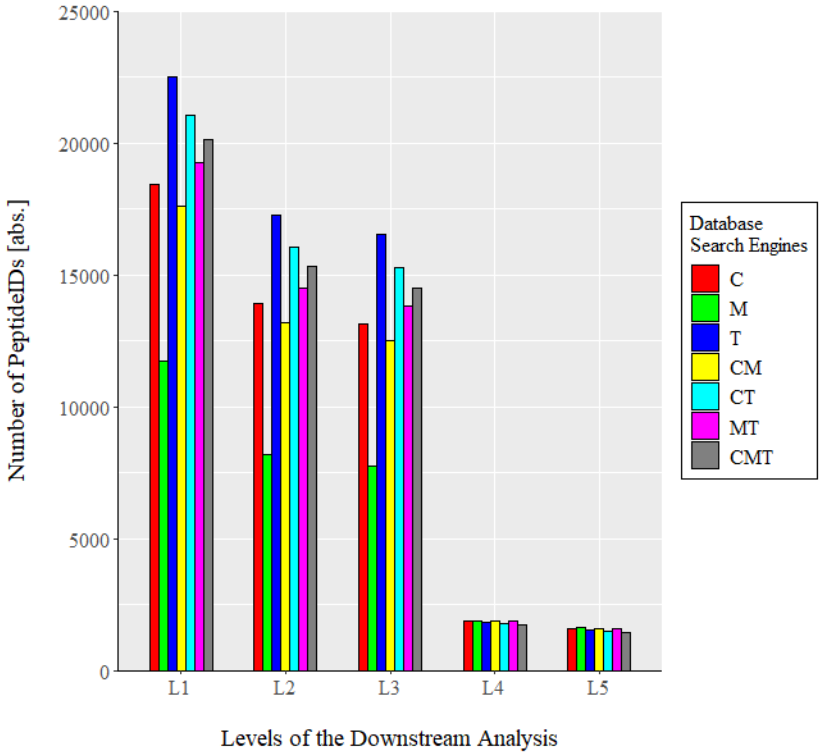


Fig. 34: Development of the number of PeptideIDs [abs.] during downstream analysis for the different database search engine combinations.

The development of the performance-based ranking on peptide-level, which is presented in Fig. 35, reflects similar behavior as on protein-level. In detail, on peptide-level the order stays constant from Level 1 to Level 3. The transition from Level 3 to Level 4 changes the ranking considerably. Best performing combinations such as T and CT decline to lower ranks. In contrast, MT and M improve their performance from rank 6 and 7 on Level 1 to rank 1 and 2 on Level 5, respectively. Moreover, each other combination also changes the rank from the filter step Level 3 to Level 4.

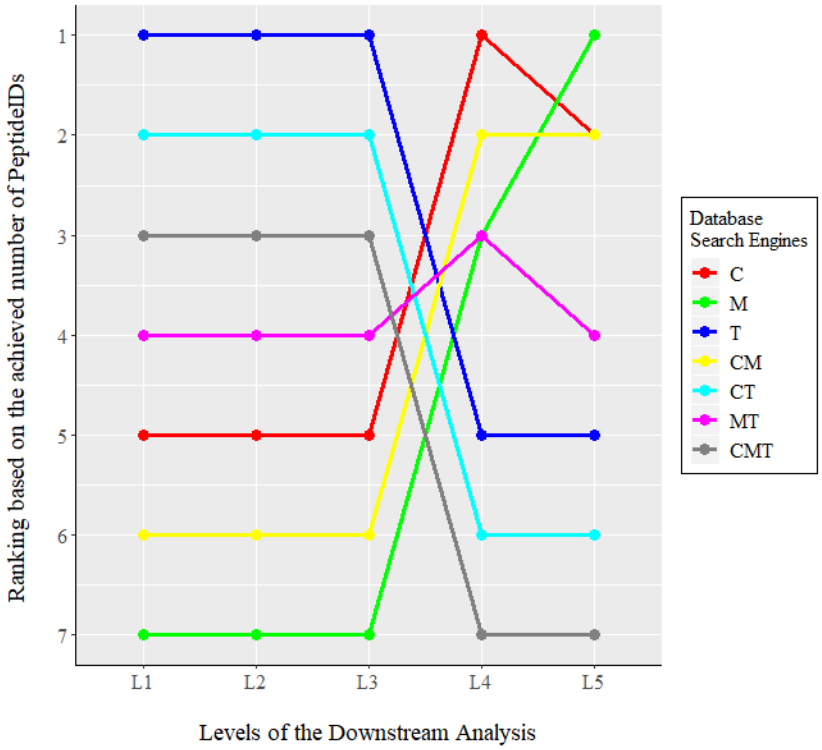


Fig. 35: Ranking based on the achieved number of PeptideIDs for the different database search engine combinations during downstream analysis. Libraries generated with an FDR < 1%.

8.1.3.3 Analysis of the influence of signal intensity and retention time variation on SWATH quantification performance

Previous findings show that in the transition from Level 3 to Level 4 nearly 90% of the results are excluded. For further evaluation of the quality of the DIA data extraction, the assigned signal intensity of the transitions stored in the library, as well as the CV of the retention time of transitions are compared between Level 3 and Level 4. In short, the DIA data is imported on Level 3 and then low-quality data (dotP < 0.8) is excluded leading to Level 4. At Level 3 and at Level 4, for each assignment the signal intensities of the corresponding transitions in the library are extracted, respectively. Next, the extracted transition signal intensities are averaged per precursor.

The corresponding averaged transition signal intensities extracted from the library, which have a protein assignment on Level 3 are displayed in Fig. 36. Subsequently, the averaged transition signal intensities extracted from the library, which have a protein assignment on Level 4 are shown in Fig. 37. Both comparisons are displayed via boxplots. Note, that outliers are not presented to achieve a better overview.

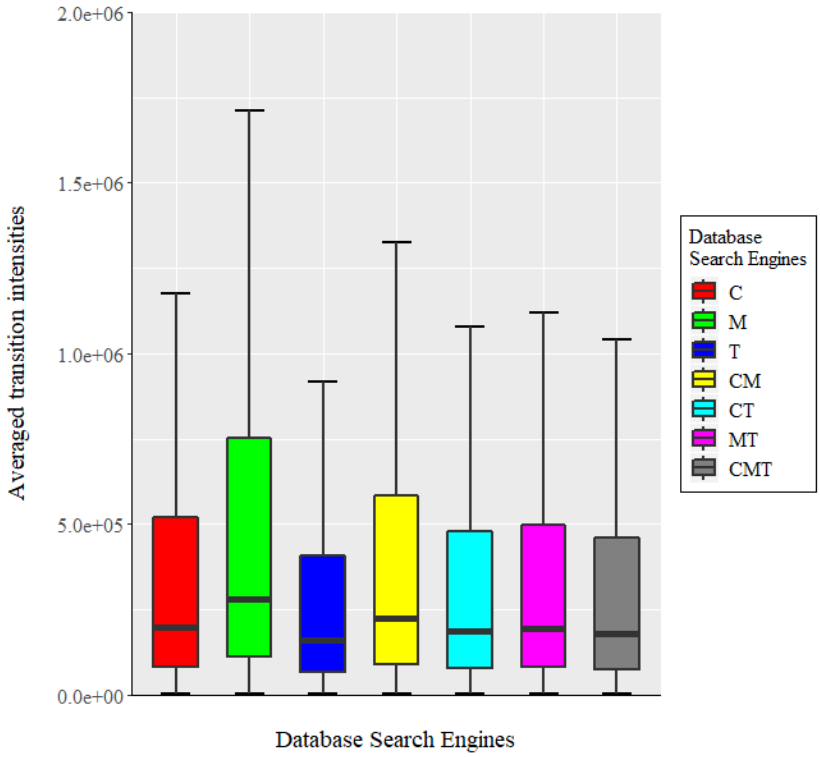


Fig. 36: Averaged transition intensities for different database search engine combinations, which are stored in the library and assigned to proteins on data analysis Level 3. Outliers are not shown.

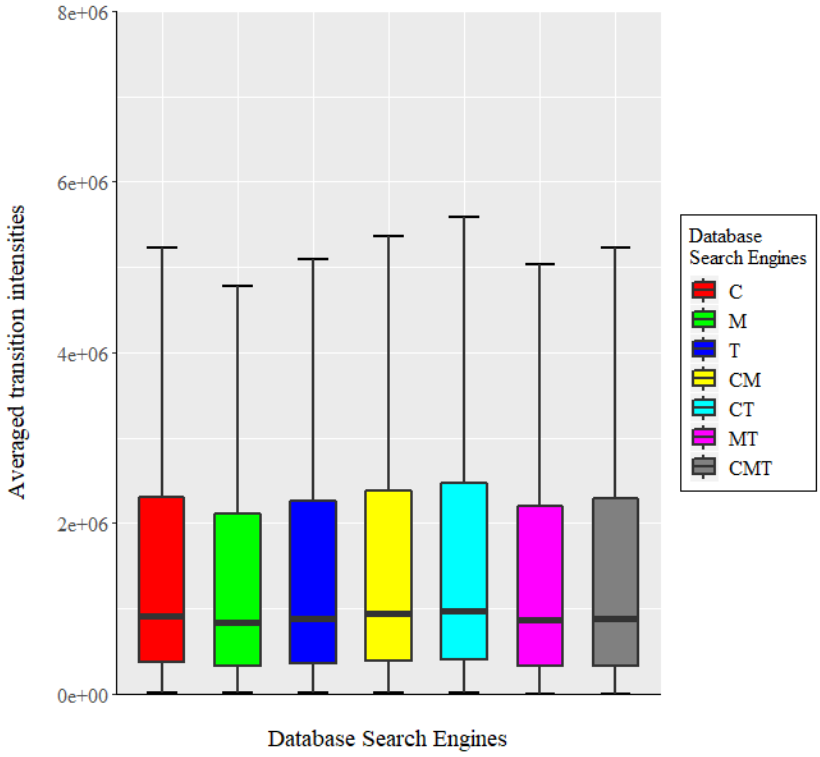


Fig. 37: Averaged transition intensities for different database search engine combinations, which are stored in the library and assigned to proteins on data analysis Level 4. Outliers are not shown.

In Fig. 36 the median of the averaged library intensities assigned to proteins on Level 3 lies within the range 1.59e+05 and 2.78e+05 across the database search tools. In comparison, the transitions of the different database search tools, which are stored in the library and assigned on Level 4, have a median signal intensity between 8.34e+05 and 9.65e+05 (Fig. 37). The results indicate that mainly precursors with high signal intensities achieve a dotP > 0.8 and thus correlate with a high-quality DIA data extraction.

Moreover, the CV of retention times for each transition was extracted and averaged per precursor for each database search engine combination. The outcomes for Level 3 are shown in Fig. 38 and for Level 4 in Fig. 39.

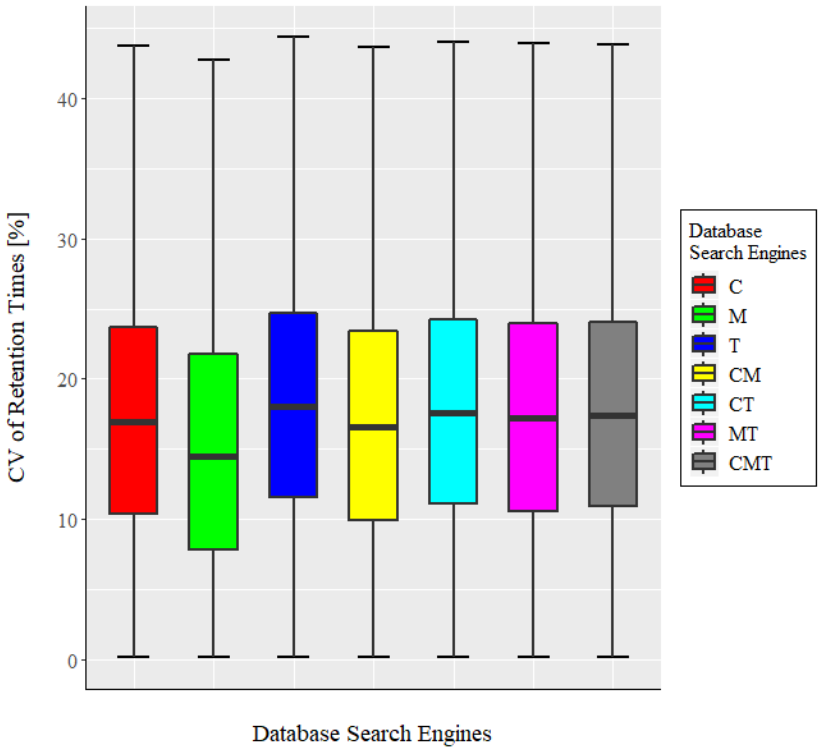


Fig. 38: CV of retention times [%] per precursor for different database search engine combinations on Level 3. Outliers are not shown.

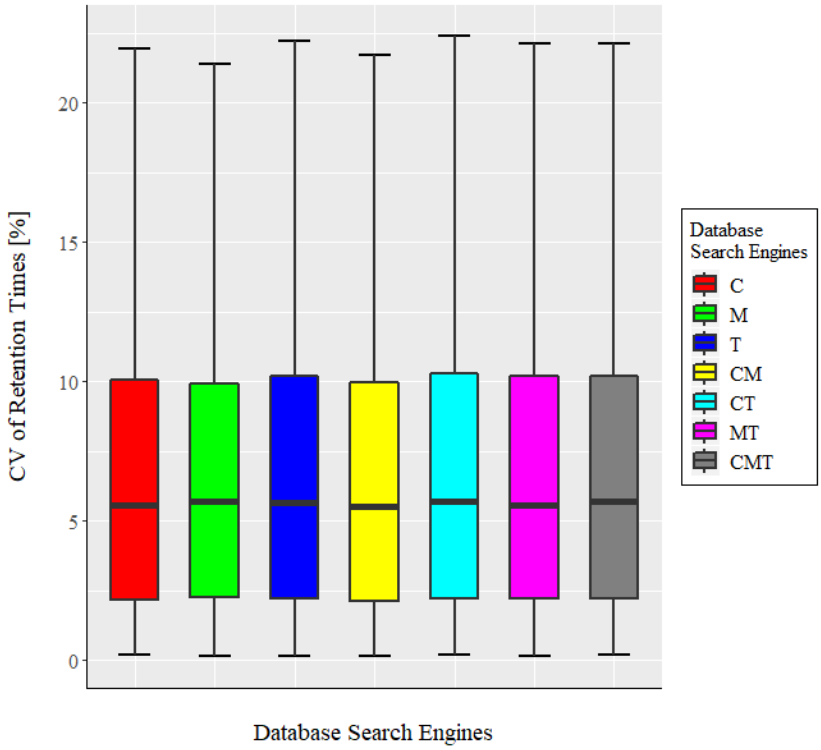


Fig. 39: CV of retention times [%] per precursor for different database search engine combinations on Level 4. Outliers are not shown.

The median of the CV of the retention times on Level 3 is leveling off at 18% for each database search engine expanding to around 43% at the upper hinge of the boxplot (Fig. 38). In contrast, the median of the CV on Level 4 is approximately 6% across every database base search combination with values of the upper hinge of the boxplot around 23% (Fig. 39). The comparison between Level 3 and Level 4 suggests that especially high CVs of retention times are prone to low-quality DIA data extraction.

In total, the results indicate that transitions with high signal intensities and low retention time variability across samples correlate with an efficient SWATH quantification performance.

8.1.4 Analysis of the consistency of detecting statistically significant proteins

In the following chapter, the consistency of determining statistically significant proteins across all database search engine combinations is investigated. In detail, the total number of statistically significant proteins in stage-wise comparisons and the similarity between the results is examined. A detailed view via volcano plots for each stage-wise comparison for the

respective search tools is presented in the appendix (Fig. 80 - Fig. 103). Proteins with an FDR of max. 5% are considered as statistically significant. Note, that biological inference with concentration on the identity of individual proteins and possible biological patterns is performed in “Part V – Biological Inference”.

Since the number of protein identifications differs after DIA analysis (“Level 5 – Refined”; Fig. 13) for the different search tool combinations, an overview for the detected proteins is presented in Fig. 40. Note, that the number of statistically analyzed proteins is relatively low because of a poor SWATH quantification performance as presented in the previous chapter. However, the binary combination MT achieves 334 protein identifications. Next in the ranking is C with 328 findings, closely followed by CM. The single search tools M and T obtain similar results with 322 and 320, respectively. Last rank CT with 312 and CMT with 306 protein identifications.

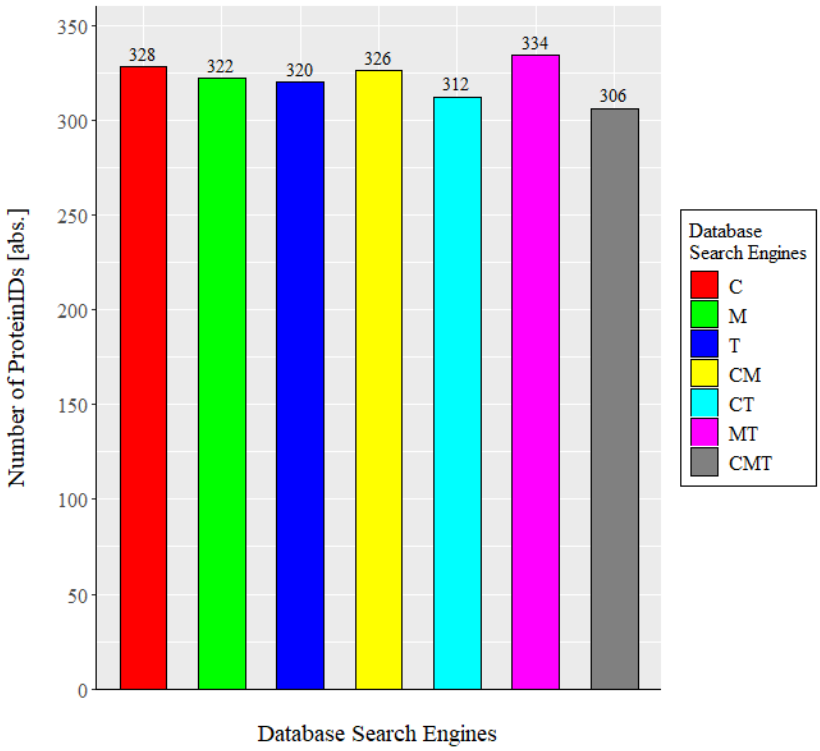


Fig. 40: Number of proteins [abs.] after Skyline analysis for the different database search engine combinations.

The coverage of proteins after the downstream analysis in Skyline is shown in Fig. 41 to study the similarity of the findings in detail. The ranking behaves according to the absolute numbers

of proteins with MT performing best with 81% and CMT ranking last with 75%. However, the overall coverage levels off around 80% and no database search engine combination covers all potentially present proteins at Level 5.

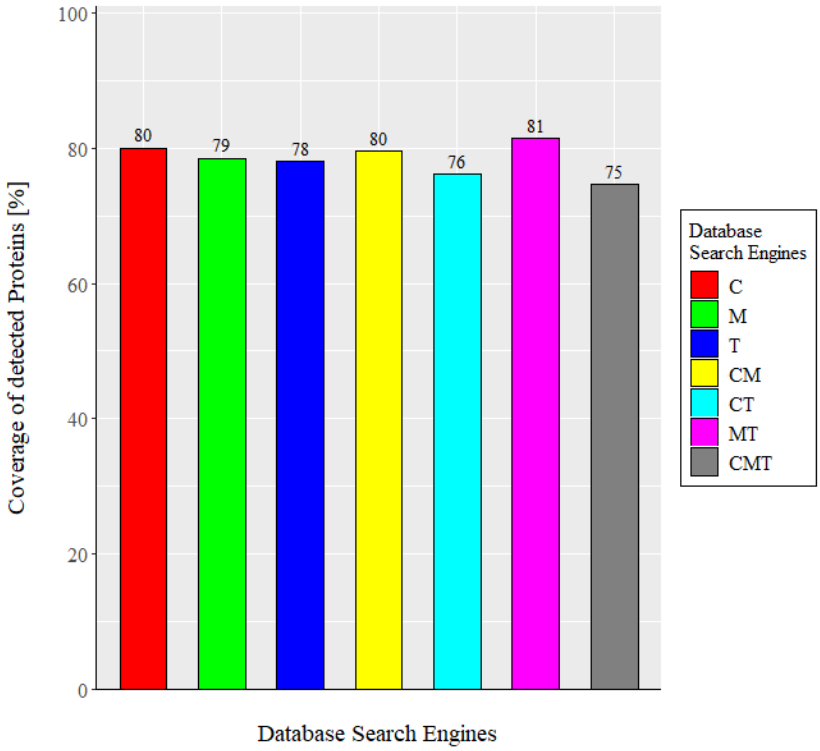


Fig. 41: Coverage of proteins [%] after Skyline analysis for the different database search engine combinations.

A detailed overview of the stage-wise comparisons, which show the number of the statistically significant hits and the resulting coverage for a specific search tool, are presented in table 3. Note, that statistically significant hits were only identified for the comparisons SI vs. SIV, SII vs. SIII, as well as SII vs. SIV. Furthermore, the total number of statistically significant hits for a given database search engine is determined (displayed as “Total” in table 3).

Table 3: Number of statistically significant hits (FDR < 5%) and coverage for stage-wise comparisons and the respective database search engine.

Database search engines	**SI vs. SIV** ProteinIDs [abs.]/ Coverage [%]	**SII vs. SIII** ProteinIDs [abs.]/ Coverage [%]	**SII vs. SIV** ProteinIDs [abs.]/ Coverage [%]	**Total** [a)] ProteinIDs [abs.]/ Coverage [%]
C	5/38	37/47	2/67	42/48
M	9/69	58/74	0/0	65/74
T	3/23	41/53	0/0	44/50
CM	5/38	52/67	0/0	56/64
CT	6/46	51/65	2/67	57/65
MT	5/38	27/35	0/0	32/36
CMT	11/85	36/46	0/0	45/51

[a)] "Total" refers to the combination of statistically significant protein of each stage-wise comparison for a respective database search engine excluding duplicate entries.

In the comparison SI vs. SIV the search tool combination CMT performs best with 11 significant hits corresponding to a coverage of 85%. Least detections are obtained by T with 3 hits. Therefore, the range of coverage for individual database search engines is within 23% and 85%. This high discrepancy in the identification rate of statistically significant proteins is also shown in the comparison SII vs. SIII. Best performing search engine is M with 58 hits corresponding to 74% coverage. In contrast, the search tool combination MT obtains 27 significant findings, which correlates with a coverage of 35%. Interestingly, for the comparison SII vs. SIV only C and CT detect significant hits. In detail, both identify 2 statistically significant proteins, which corresponds in both cases to a coverage of 67%. Considering the sum of the statistically significant hits, the search tool M performs best with 65 hits and the binary combinations MT worse with 32 findings.

In total, the results of the stage-wise comparisons show partially major differences in detecting statistically significant proteins. No search tool is able to identify all statistically significant hits. The question about why a certain database search engine combination performs better than the other and if there is a possible connectivity between the library size and the detection of statistically significant proteins will be discussed in the next chapter.

8.2 Discussion

In this chapter, the previous shown results in terms of library size, analysis time, data storage size, downstream analysis and statistically significant proteins will be further examined and evaluated. In addition, potential correlations between categories will be studied to highlight potential dependencies and investigate advantages of combining the results of multiple database search engines.

8.2.1 Library size, analysis time, and storage size

The analysis of the library size shows that a single search engine can obtain a higher library input than a binary combination, as well as a binary combination can outperform the triple combination CMT. Hence, merging the outcomes of multiple search engines does not necessarily enhances the library size. The performance on peptide- and precursor-level are influenced among other factors mainly by the performance as single database search tool and the complementarity between database search engines within combinations. As previously stated, the scoring functions of the search tools C and T are based on a descriptive model. In contrast, the search engine M applies an interpretative method. Since the overlap between sets of identified spectra by search engines is known to be less, if the search tools use different scoring strategies, CM and MT should outperform CT. However, the single search tool T excels all other options. As a result, CT performs better than CM and MT. For the single variants M and C the performance is always enhanced by adding results of one further search engine. Moreover, it is noteworthy that the determined performances of the database search engines are primarily limited to the present dataset.

Regarding the analysis time and storage size the single search tool M ranks first, closely followed by T. Moreover, combining the outcomes of search engines always results in a significant increase in analysis time. In detail, starting from a single search engine, the analysis time increases up to 65% to a binary option and around 100% to the triple combination. In addition, it is obvious that joining results of search engines to a specific search tool invariably leads to an increased memory requirement. While the triple combination needs around 95 GB, the single search tool T only requires 52 GB. Both the high analysis time and file storage size are mainly based on using all 70 DIA files for the data analysis. Additional investigation on the impact of the amount of DIA files on the library size and on further data analysis is desirable

because it may reduce the computational costs. However, the single search tool T excels in terms of the library input all other options. In addition, since merging results increases the analysis time and memory space significantly, T is considered as valuable choice for this dataset within the DDA-free analysis.

8.2.2 SWATH quantification performance and reproducibility of the detection of statistically significant proteins

By performing the downstream analysis only 10% of high-quality identifications remain. Especially the DIA extraction and the subsequent removal of low-quality assignments result into a decline of nearly 90%. Further data mining suggests that high-abundant transitions and a low retention time variability are key characteristics to achieve a high-quality DIA data extraction and ensure an efficient SWATH extraction performance. In addition, the performance-based ranking of the different database search engine combinations is altered by this transition. For example, the single search tool T performs best on library-level and only ranks 5th on Level 5. As a conclusion, enhancing the information input on library-level does not necessarily correlate with an adequate DIA data extraction. However, further examining the influence of a specific library spectrum of a certain database search engine on a dotP for a given assignment and on the data analysis considering the whole information input of the corresponding library is beyond the scope of this project. Note, that the discussion is mainly restricted to this dataset.

The detection of statistically significant proteins varies across the database search engine combinations. Since the identification of proteins on Level 5 inherits a relatively high variation with a coverage around 80% for the database search engine combinations, the identification of statistically significant proteins is more likely to differ. Hence, the generated library and its input for a specific search tool impacts the downstream analysis and consequently the last data analysis level determines statistical inference. These correlations emphasize the importance of the library choice. Furthermore, the results of the statistical analysis show the limitations of only using a specific library. No database search engine combination and the resulting library covers the detection of all statistically significant proteins. Another advantage of performing an analysis with more than one library is the fact that detection of the same statistically significant proteins, enhances confidence and verifies the results.

8.3 Conclusion

The objective of the DDA-free analysis was to enhance the identification rate on library-level by combining the results of multiple database search engines for library generation to extract the information of DIA data without the necessity of acquiring DDA data. An overview of advantages and disadvantages of each database search engine combination is shown in table 4.

In most of the cases merging the results of one specific search tool to a second tool enhanced the identification rate on library-level. Only the single search engine T excelled. In this case, adding results from another engine had no further benefits for library generation. Since for the library generation all 70 DIA files were used, combining results had a significant impact on the analysis time and storage size. Further investigation on how many DIA files are necessary to achieve similar results may reduce the computational costs. In addition, the DIA analysis showed that the increase of the library information is not proportional to SWATH quantification performance. The quality of library input regarding signal intensity and the retention time variability of the transitions played a key role. Moreover, the choice of library directly impacted the detection of statistically significant proteins. No database search engine combination covered the identification of all possible statistically significant proteins (FDR < 5%).

Table 4: Overview of merits and drawbacks of each database search engine combination.

Categories	C	M	T	CM	CT	MT	CMT
Analysis time	+	++	++	-	-	-	--
Storage size	+	++	++	-	-	+	--
Number of precursors in Library (FDR < 1%)	-	--	++	-	++	+	++
Number of peptides in Library (FDR < 1%)	-	--	++	-	++	+	++
Number of extracted proteins (FDR < 1%) [a)]	+	-	-	+	-	++	--
Extraction of statistically significant proteins	-	++	-	+	+	--	-

[a)] Refers to Level 5 – after all filtering steps are performed (see Fig. 13).

While there is no warranty for an increase on library size by combining multiple search engines, it is still recommendable to use the results of two database search engines and run the total analysis both for the search tools individually and their combination in order to exploit the valuable information provided by the DIA data. It not only adds certainty to the determined results but also increases the chance to identify statistically significant proteins and highlight potential pathogenic patterns for CRC.

9. Part IV – Comparison: DDA-based vs. DDA-free Analysis

9.1 Results

In this chapter, the two applied data analysis strategies DDA-based and DDA-free will be compared as a whole regarding library size, computational costs, SWATH quantification performance, and the extraction of statistically significant proteins.

9.1.1 Library size

The library size for the respective data analysis strategy is examined on precursor- and peptide-level (Fig. 42 and Fig. 43). To elaborate, quartiles are calculated based on the achieved number of identifications across each database search tool at an FDR < 1% for the corresponding data analysis strategy and then visualized via boxplots.

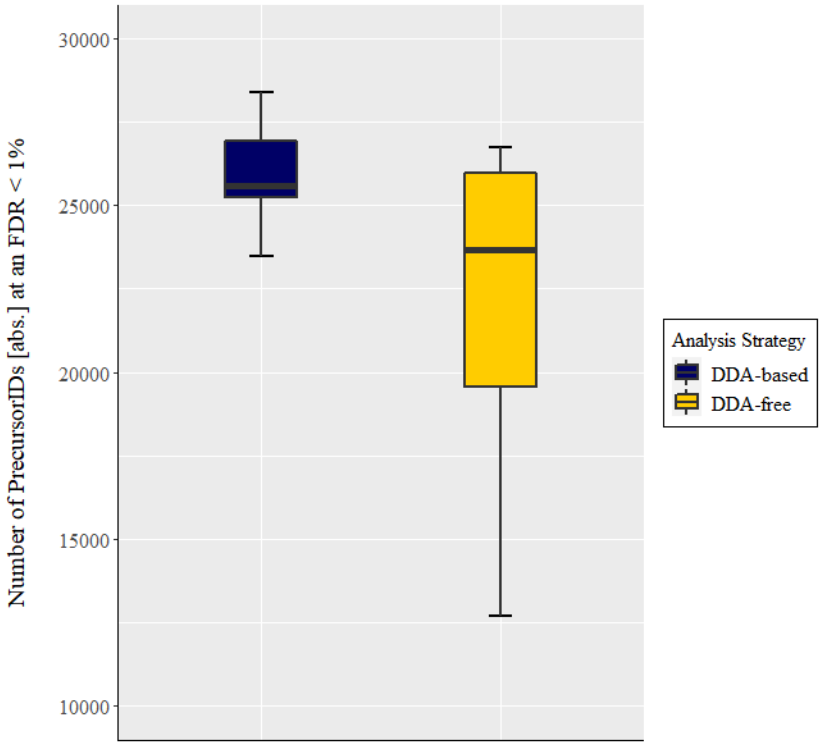


Fig. 42: Library Size – number of precursors [abs.] at an FDR < 1% for different data analysis strategies: DDA-based vs. DDA-free.

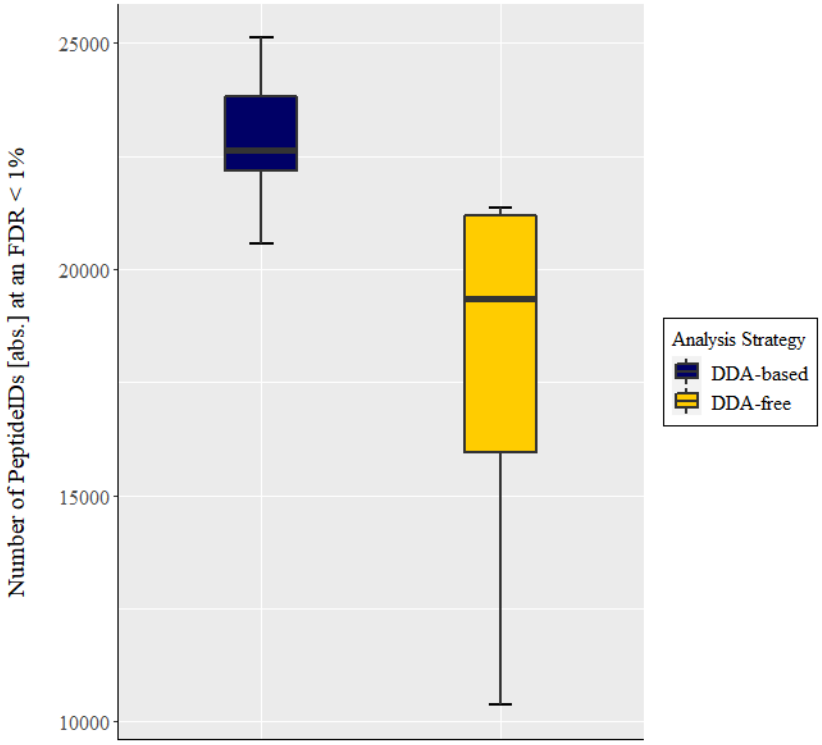


Fig. 43: Library Size – number of peptides [abs.] at an FDR < 1% for different data analysis strategies: DDA-based vs. DDA-free.

Both on precursor- and on peptide-level the DDA-based approach obtains a higher library input. The median number of precursor identifications for the DDA-based strategy is 25588 in comparison with 23668 for the DDA-free data analysis (Fig. 42). On peptide-level the DDA-based data analysis obtains a median of 22642 hits and the DDA-free strategy 19360 identifications (Fig. 43). Additionally, both Fig. 42 and Fig. 43 show a higher variability for the DDA-free approach.

9.1.2 Analysis time and file storage size

The computational costs of the different data analysis approaches are evaluated regarding analysis time and file storage size (Fig. 44 and Fig. 45). Both characteristics correspond to a data analysis with an FDR < 1%. For the obtained values by different database search combinations quartiles are computed per data analysis strategy and displayed via boxplots, respectively.

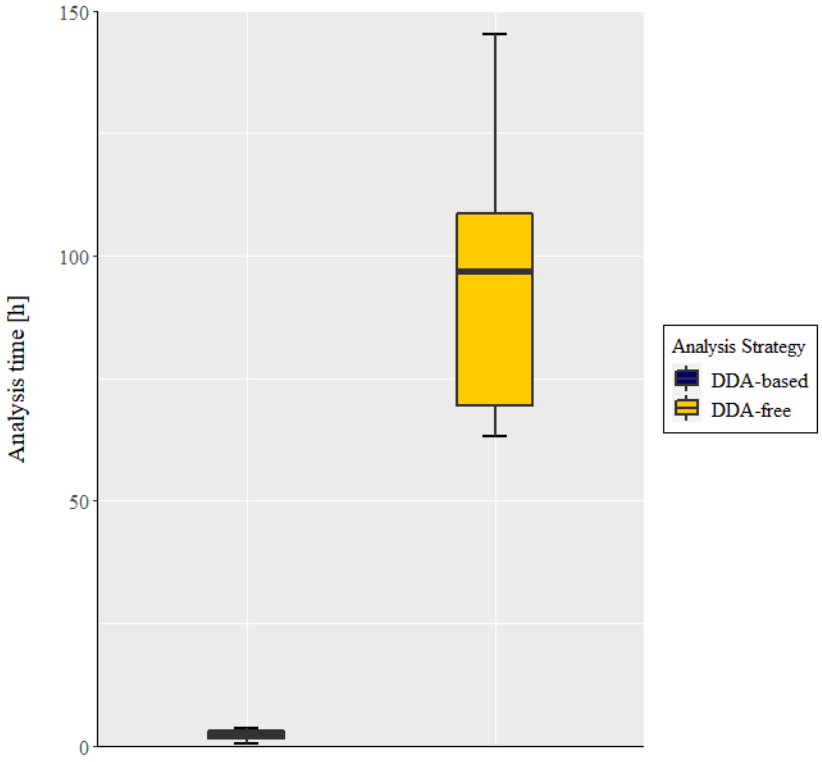


Fig. 44: Analysis time [h] for different data analysis strategies: DDA-based vs. DDA-free.

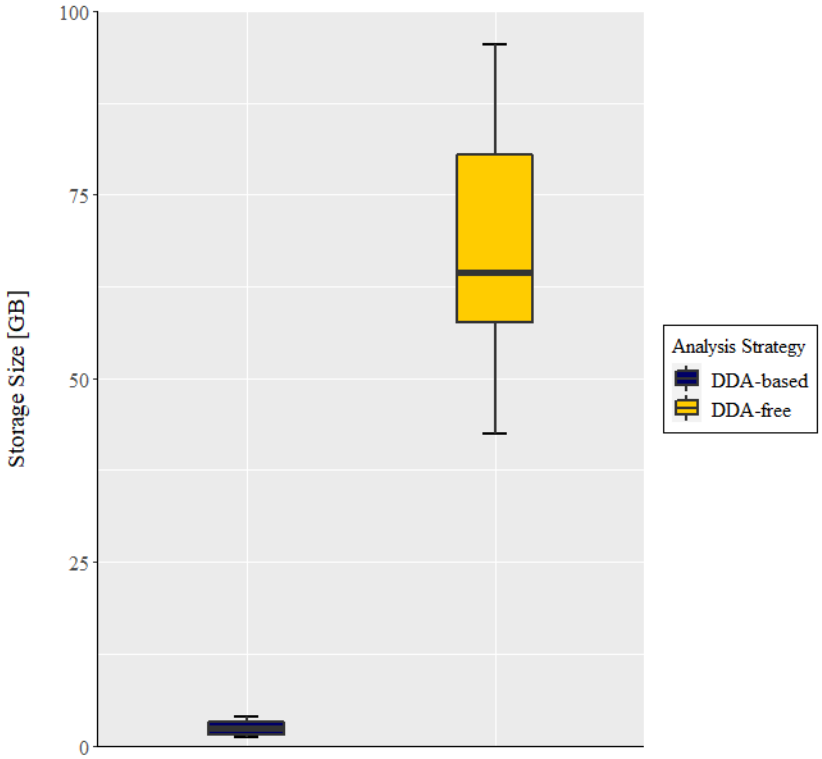


Fig. 45: File storage size [GB] for different data analysis strategies: DDA-based vs. DDA-free.

The median data analysis time of the DDA-based strategy is significantly lower with 2.11 h than then the needed time for the DDA-free approach with 96.8 h (Fig. 44). In addition, the median memory requirement of the files of the DDA-based data analysis with 2.36 GB outperforms the DDA-free path with 64.4 GB (Fig. 45). Moreover, the DDA-free strategy depicts a higher variability both in the analysis of the time and storage size.

9.1.3 SWATH quantification performance on protein- and peptide-level

The SWATH quantification performance of the different data analysis approaches is evaluated by comparing Level 3 and Level 4 on protein- and peptide-level. This transition corresponds to a removal of low-quality assignments with a dotP < 0.8 (for details see Fig. 13). On both Level 3 and Level 4 the obtained number of identifications by the different database search engine combinations is used to calculate quartiles per data analysis strategy. The results on protein-level are displayed via boxplots for Level 3 in Fig. 46 and for Level 4 in Fig. 47. The outcomes on peptide-level are shown in Fig. 48 and Fig. 49, respectively.

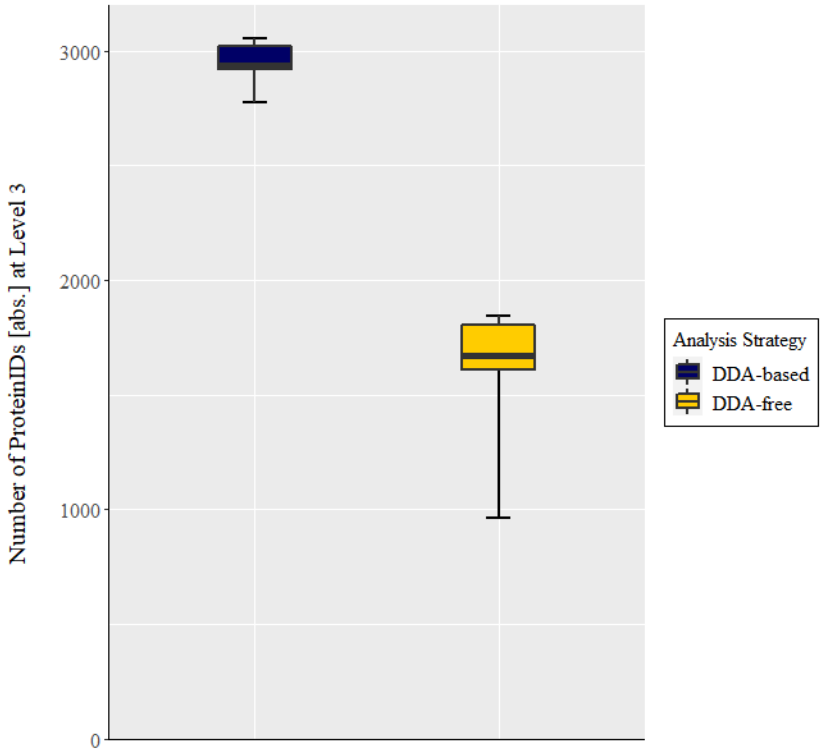


Fig. 46: Number of ProteinIDs [abs.] at Level 3 for different data analysis strategies: DDA-based vs. DDA-free.

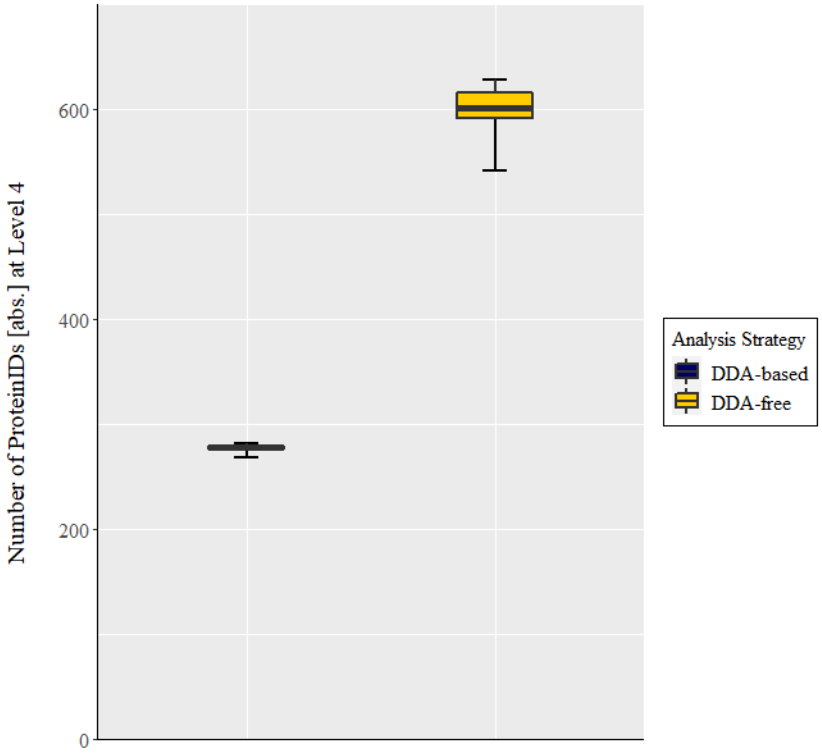


Fig. 47: Number of ProteinIDs [abs.] at Level 4 for different data analysis strategies: DDA-based vs. DDA-free.

The DDA-based path achieves a median of 2938 protein identifications on Level 3 in comparison to 1672 identifications for the DDA-free analysis (Fig. 46). On Level 4 for the DDA-based strategy 278 high-quality assignments and for the DDA-free approach 601 protein identifications remain (Fig. 47).

Similar behavior is displayed on peptide-level (Fig. 48 and Fig. 49). While the DDA-based strategy achieves more peptide identifications on Level 3 with a median of 17631 than the DDA-free path with 13835, the performance changes on Level 4. The DDA-free approach obtains a median of 1868 high-quality assignments and the DDA-based analysis achieves 547 peptide identifications.

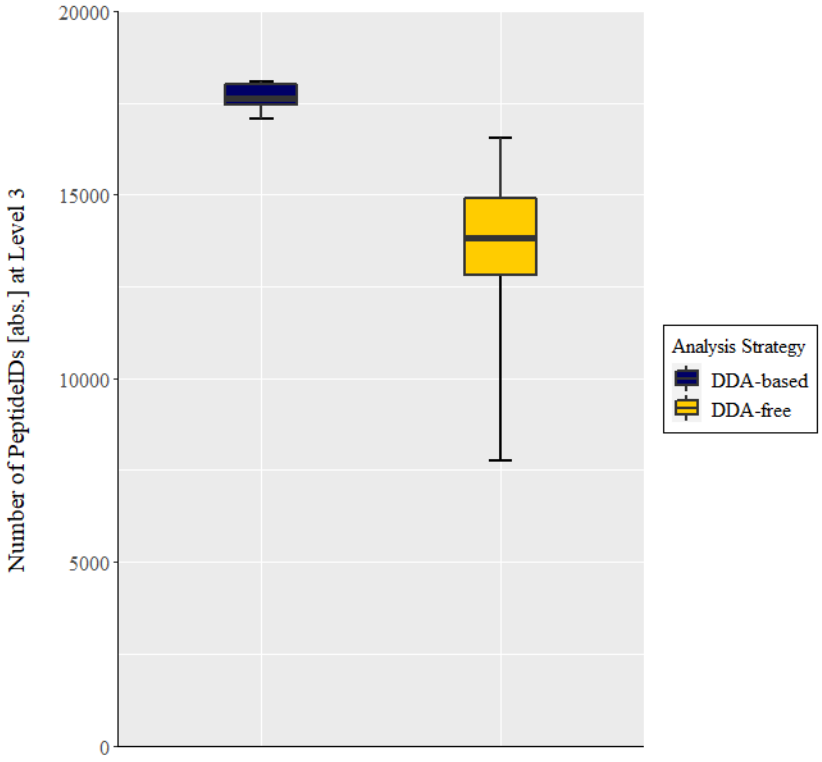


Fig. 48: Number of PeptideIDs [abs.] at data analysis Level 3 for different data analysis strategies: DDA-based vs. DDA-free.

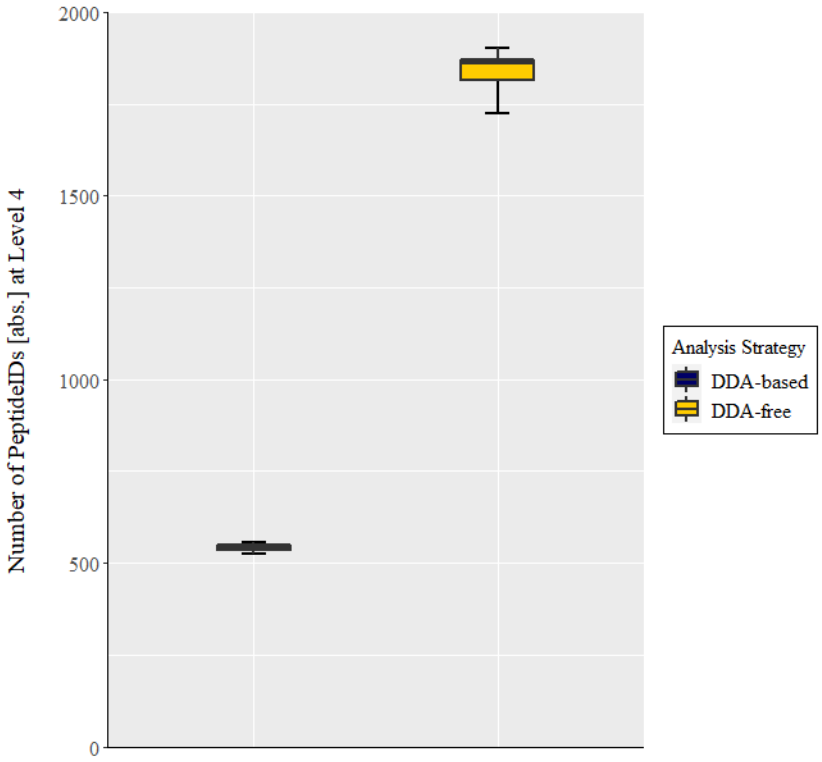


Fig. 49: Number of PeptideIDs [abs.] at data analysis Level 4 for different data analysis strategies: DDA-based vs. DDA-free.

9.1.4 Analysis of the influence of signal intensity and retention time variation on SWATH quantification performance

To further investigate the differences in the SWATH quantification performance for the data analysis strategies the influence of the signal intensities in the library and the CV of the retention times on the downstream analysis are examined.

For each data analysis approach the DIA data is imported on Level 3 and subsequently low-quality data (dotP < 0.8) is removed resulting into Level 4. At Level 3 and at Level 4, for each assignment the averaged signal intensities of the corresponding precursor in the library are extracted for a specific search tool. Next, quartiles are computed based on the determined precursor intensities of each database search engine per data analysis strategy. The resulting boxplots for Level 3 are shown in Fig. 50 and for Level 4 in Fig. 51.

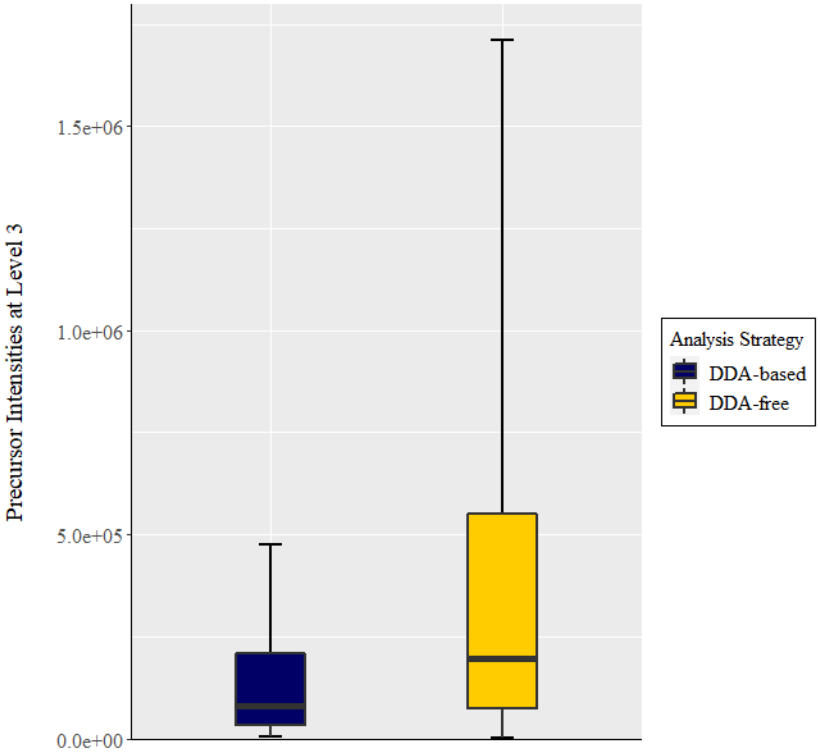


Fig. 50: Precursor intensities stored in the library for different data analysis strategies at data analysis Level 3: DDA-based vs. DDA-free.

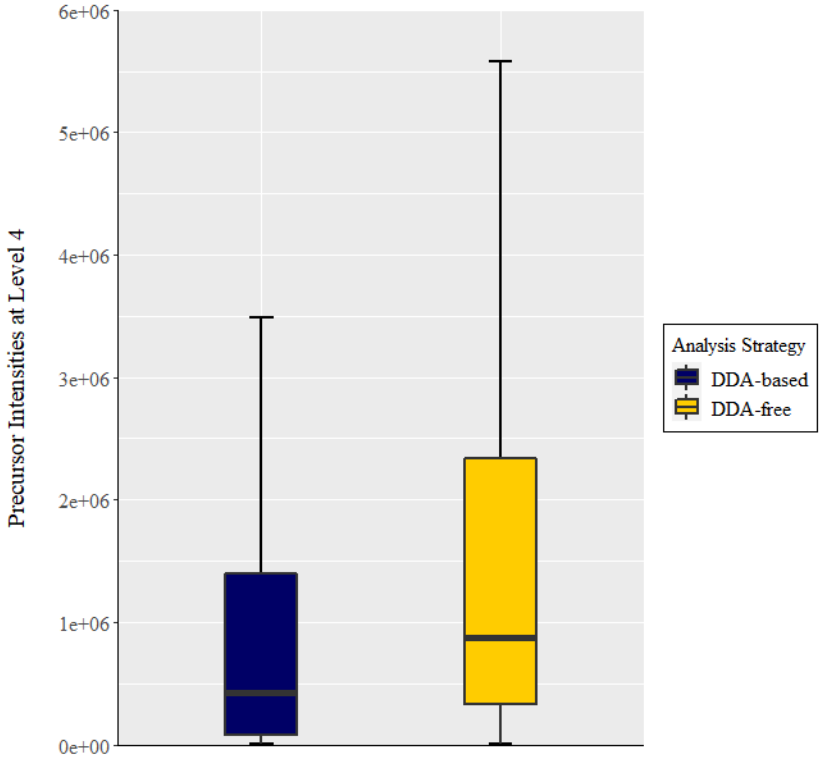


Fig. 51: Precursor intensities stored in the library for different data analysis strategies at data analysis Level 4: DDA-based vs. DDA-free.

In Fig. 50 the precursor intensity for Level 3 is presented, which includes a median precursor intensity for the DDA-based strategy of 8.11e+04 and 1.95e+05 for the DDA-free path. While both analysis strategies obtain higher intensity values on Level 4 (Fig. 51), the DDA-free strategy has again a higher median signal intensity of 8.75e+05 in comparison with the DDA-based approach with a median signal intensity of 4.19e+05.

Furthermore, the CV of retention times of the precursors for each data analysis strategy is presented in Fig. 52 for Level 3 and in Fig. 53 for Level 4. The results of each database search engines per analysis strategy is used to calculate quartiles, which are subsequently visualized via boxplots.

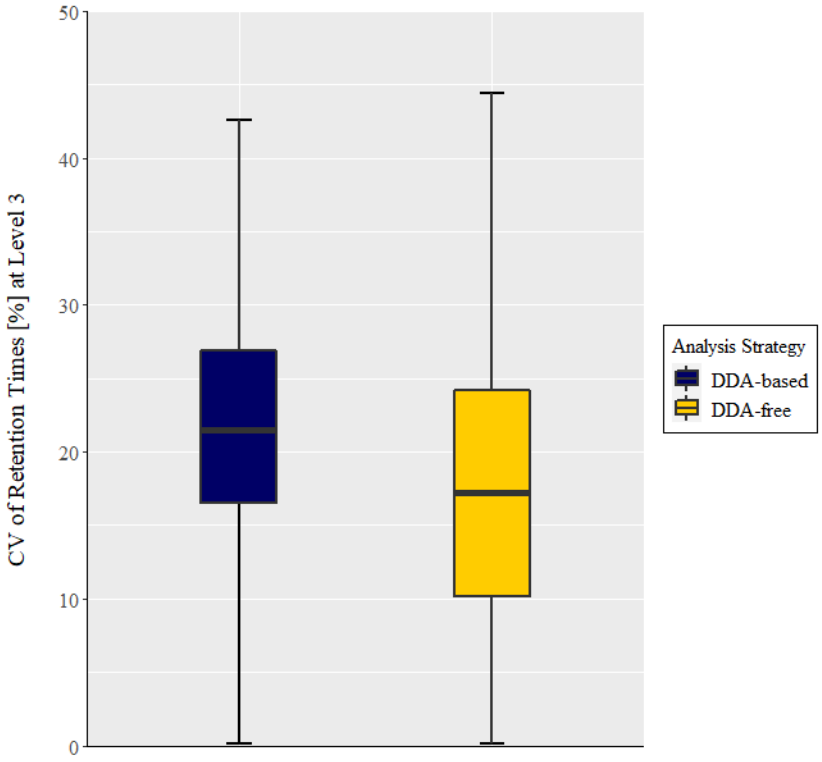


Fig. 52: CV of retention times [%] at data analysis Level 3 for different data analysis strategies: DDA-based vs. DDA-free.

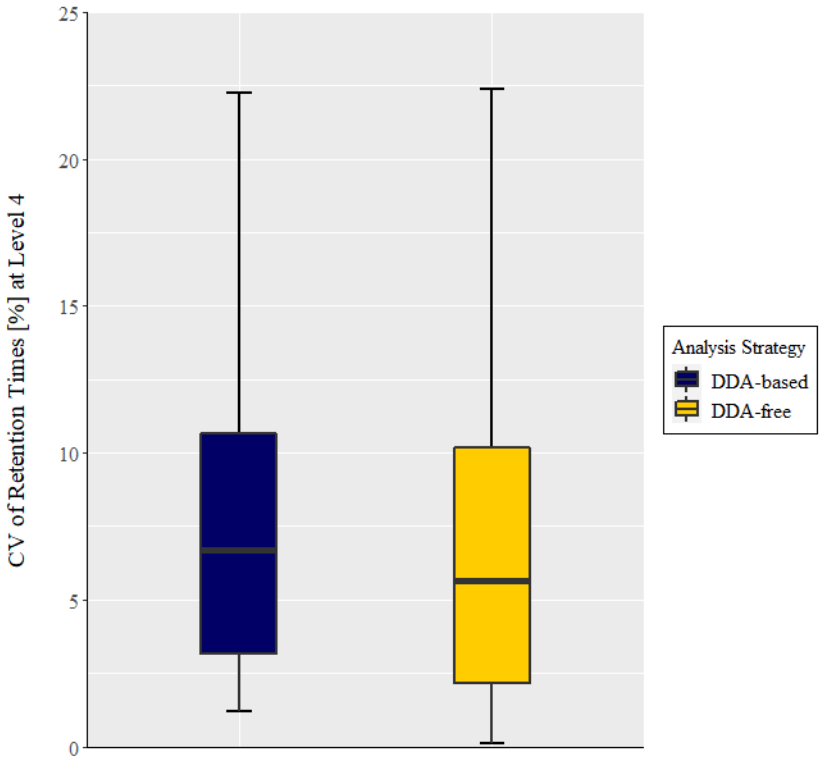


Fig. 53: CV of retention times [%] at data analysis Level 4 for different data analysis strategies: DDA-based vs. DDA-free.

On Level 3 the DDA-based analysis path displays a median CV around 21% and the DDA-free strategy about 17% (Fig. 52). For both approaches the CV decreases on Level 4. In detail, the DDA-based median CV declines to 7% and for the DDA-free approach to 6% (Fig. 53).

9.1.5 Extraction of statistically significant proteins

In this chapter, both the DDA-based and DDA-free data analysis will be compared regarding the detection of significant outcomes. In addition, the similarity of the findings will be investigated.

First, the total number of statistically significant proteins is determined per data analysis strategy (Fig. 54). To clarify, the total amount of significant findings per database search engines is used to calculate quartiles for the respective data analysis approach. The results are presented via boxplots. Note, that proteins are considered statistically significant with an FDR threshold of max. 5%.

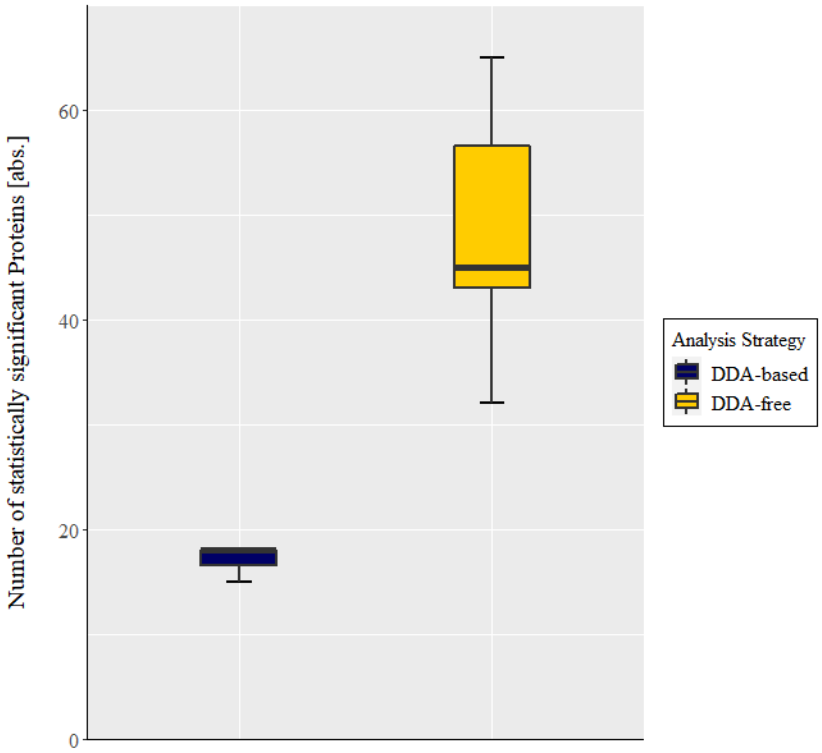


Fig. 54: Number of statistically significant proteins [abs.] for different data analysis strategies: DDA-based vs. DDA-free.

The DDA-based strategy achieves a median number of 18 statistically significant proteins and the DDA-free approach obtains a median of 45 significant hits (Fig. 54). Furthermore, the DDA-free approach varies more in terms of detecting significant findings.

Next, the similarity between the different data analysis strategies is examined (Fig. 55). To elaborate, the DDA-based strategy obtains 22 unique proteins (see table 1) and the DDA-free analysis achieves 88 unique hits (see table 3). In the next step, these identifications are compared. The Venn diagram displays 7 (7.4%) specific identifications for the DDA-based path and 73 (76.8%) for the DDA-free approach. Moreover, 15 (15.8%) statistically significant proteins are detected by both data analysis strategies.

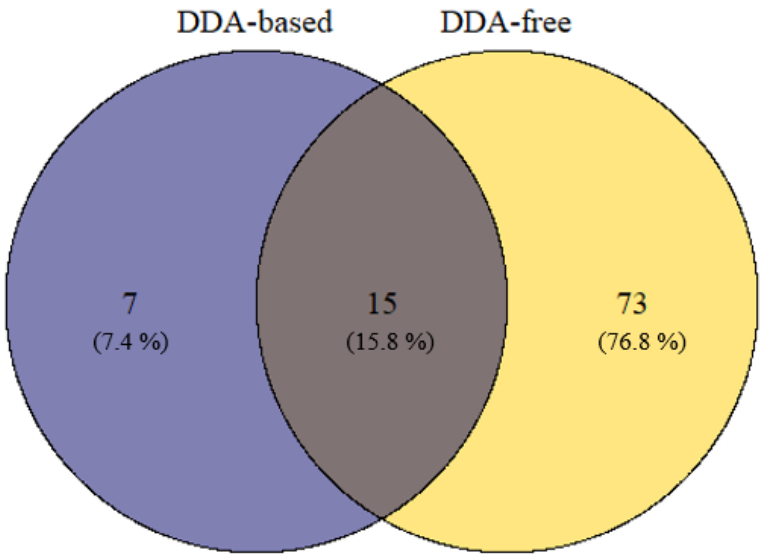


Fig. 55: Venn Diagram of the statistically significant proteins corresponding to different data analysis strategies: DDA-based vs. DDA-free.

9.2 Discussion

In the following chapter, all previously demonstrated results regarding library size, analysis time, data storage size, and extraction of statistically significant hits will be evaluated and individual characteristics will be correlated in order to highlight possible dependencies and examine potential benefits of a specific data analysis strategy.

On the one hand, the DDA-based data analysis strategy outperforms the DDA-free path significantly on peptide- and precursor-level regarding the library size. As an example, if DDA data is used for library generation around 3000 more peptides are stored in the library in comparison with the DDA-free approach. On the other hand, the SWATH quantification performance for the DDA-free strategy is significantly better than for the DDA-based approach. Note, that the SWATH quantification performance of both strategies is low, only that the DDA-free strategy performs better relative to the DDA-based path. However, the CV of retention times across samples is lower and the stored precursor intensities are higher for the DDA-free strategy. In other terms, while the overall library input of the DDA-free approach is lower, the quality of the input is higher in comparison with the DDA-based strategy. Potential reasons stem from the fact that the DIA-Umpire module performs a signal processing algorithm, which calculates for each monoisotopic peak of a precursor and fragment peak a Pearson correlation primarily based on LC elution peaks and retention times to build precursor-fragment groups. These co-eluting precursor and fragment ions form pseudo-tandem MS/MS spectra, which are subsequently used for database search (see chapter 3.5.2.2). Hence, the constructed pseudo-MS/MS spectra might include information, which resemble the acquired DIA spectra of the samples in a better way. However, in addition, the DDA-free strategy obtains more statistically significant proteins. In detail, only 7.4% of all possible statistically significant hits are not identified by the DDA-free approach. Again, it is noteworthy that all made considerations are mainly based on the applied dataset.

Another important aspect are the significant differences regarding the computational costs. In both analysis time and file storage requirement the DDA-based strategy excels the DDA-free approach. As previously discussed in chapter 8.2, the vast computational costs of the DDA-free path are based on performing an extra analysis step with DIA-Umpire and using all 70 DIA files for the pipeline. In contrast, for the DDA-based strategy 26 DDA-files are utilized. However, further optimization of the analysis time and file requirements for the DDA-free

strategy might reduce the computational costs and thus minimize the differences between both data analysis strategies. Additionally, it is important to mention, that for the DDA-based approach experimental costs including sample amount, measurement time, chemicals, fractionation procedure etc. are significantly higher in comparison to the DDA-free strategy, which is experimentally only based on the DIA measurements and the respective sample preparation.

9.3 Conclusion

The aim of both data analysis strategies was to increase the library input for subsequently extracting the information of the DIA data in a best possible way.

An overview of advantages and disadvantages of each data analysis strategy, in which both approaches are compared relative to each other, is presented in table 5. The DDA-based strategy outperforms the DDA-free path in terms of library size. In contrast, the DDA-free approach achieves a better SWATH quantification performance and extracts more statistically significant proteins. In other terms, the total library size is smaller, but the quality of the input is higher for the DDA-free approach in comparison with the DDA-based strategy. In this context, for both strategies a key characteristic is retention time variability across runs, as well as the signal intensity of the transitions. Further post-measurement optimization regarding retention time alignment might improve the SWATH quantification performance including common internal retention time standards (CiRTs) or DIAlignR [94,95]. However, in addition, experimental and computational costs differ between the two data analysis strategies. While the DDA-free approach obtains considerably higher computational costs, experimental requirements are lower in comparison with the DDA-based strategy. Furthermore, the DDA-free strategy achieves a better extraction of valuable information of the DIA data.

Table 5: Overview of merits and drawbacks of each data analysis strategy.

Categories	DDA-based	DDA-free
Library size	+	-
Analysis time	++	--
Storage size	++	--
SWATH quantification performance	-	+
Extraction of statistically significant proteins	-	+
Experimental costs	-	+

Considering each aspect of the comparison, the DDA-free strategy is viewed as a valuable option to exploit the high information provided by the DIA data especially in a setting in which sample amounts and measurement time are limiting resources. In addition, based on applying both data analysis strategies including thorough data mining the identified findings have a high verified quality. In total 15 statistically significant proteins are identified by both strategies and are submitted to biological inference in the next chapter to evaluate the potential to serve as target proteins for further research in the area of CRC.

10. Part V – Biological Inference

10.1 Results

With the aim in mind to elucidate potential pathogenic patterns for CRC in the sense of systems biology the following chapter focuses on the biological inference of the determined statistically significant proteins. Only the findings which are verified by each data analysis strategy are directed to biological inference to provide a high degree of authenticity. An overview of the identified proteins is displayed in table 6. First, the identified proteins are subjected to a network and pathway analysis. Subsequently a literature search of the hits in the context of CRC will be performed. Both network analysis and literature mining are essential to check the plausibility of the results and elaborate their potential for further studies.

Table 6: Statistically significant proteins which are identified both by the DDA-based and DDA-free data analysis strategy.

UniProtID	Protein names	Gene names
P23396	40S ribosomal protein S3	RPS3
P05387	60S acidic ribosomal protein P2	RPLP2
P23526	Adenosylhomocysteinase	AHCY
P04083	Annexin A1	ANXA1
P23528	Cofilin-1	CFL1
P01024	Complement C3	C3
P04843	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 1	RPN1
P17931	Galectin-3	LGALS3
P68871	Hemoglobin subunit beta	HBB
P02042	Hemoglobin subunit delta	HBD
P32119	Peroxiredoxin-2	PRDX2
P30044	Peroxiredoxin-5	PRDX5
P25815	Protein S100-P	S100P
P10599	Thioredoxin	TXN
P08670	Vimentin	VIM

10.1.1 Pathway and network analysis

Network and pathway analysis are based on the reactome pathway knowledgebase. In general, the database offers molecular insights about signal transduction, transport, metabolism, and further cellular processes [96,97]. In combination with Cytoscape, a biological network and analysis platform, the ReactomeFIViz app was used for a pathway enrichment analysis [98]. The results of the enrichment analysis for biological pathways within the reactome knowledgebase are presented in table 7. Only pathways which achieve an FDR < 5% and include at least 3 hit genes are elaborated further.

The proteins cofilin-1 (*CFL1*), 40S ribosomal protein S3 (*RPS3*) and 60S acidic ribosomal protein P2 (*RPLP2*) are enriched in the pathway "Axon guidance", which is the process induced by neurons to direct axons to a specific target. In detail, a growth cone situated at the tip of axons reacts to environmental signals and responds with attractive or repulsive movements [99]. Ribosomal proteins, as well as cofilin-1, have been reported to be guidance cues for this process [99,100]. Furthermore, studies are also indicating that cancer cells are able to stimulate neuronal growth towards the tumor and thus impact tumor growth and migration [101,102].

Moreover, the proteins 40S ribosomal protein S3 (*RPS3*), Dolichyl-diphosphooligosaccharide-protein glycosyltransferase subunit 1 (*RPN1*) and 60S acidic ribosomal protein P2 (*RPLP2*) are enriched in the process "SRP-dependent cotranslational protein targeting to membrane". In general, translation refers to protein synthesis from an mRNA sequence. Contranslational targeting involves the delivery of nascent proteins while the translating ribosome is still attached [103]. For example, proteins destined for the endoplasmic reticulum (ER) are submitted to the ER by a cytosolic signal recognition particle (SRP). In brief, the corresponding polypeptide is translocated into the ER while elongation of the translation continues [104]. As an example, Dolichyl-diphosphooligosaccharide-protein glycosyltransferase subunit 1 plays a role in the translocation process [105]. While ribosomal proteins have several crucial cellular functions in the process of cotranslation, they also have been reported to be overexpressed in the context of CRC [103,106].

Table 7: Enrichment analysis for biological pathways within the reactome pathway database (FDR < 5%) [96,97].

Pathway	Proteins in Pathway	Proteins from Gene Set	FDR < 5%	Hit Genes
Metabolism of amino acids and derivatives	273	3	3.85E-02	AHCY, RPS3, RPLP2
Selenoamino acid metabolism	105	3	1.48E-02	AHCY, RPS3, RPLP2
Signaling by interleukins	436	3	4.09E-02	ANXA1, CFL1, VIM
Axon guidance	492	3	4.18E-02	CFL1, RPS3, RPLP2
Neutrophil degranulation	423	3	4.09E-02	HBB, C3, LGALS3
Innate immune system	998	5	3.85E-02	HBB, TXN, C3, LGALS3, CFL1
SRP-dependent cotranslational protein targeting to membrane	103	3	1.48E-02	RPN1, RPS3, RPLP2
RNA polymerase II transcription	885	4	4.75E-02	TXN, LGALS3, PRDX2, PRDX5
Generic transcription pathway	764	4	4.09E-02	TXN, LGALS3, PRDX2, PRDX5
Transcriptional regulation by *TP53*	339	3	3.85E-02	TXN, PRDX2, PRDX5
Cellular responses to stress	327	3	3.85E-02	TXN, PRDX2, PRDX5
TP53 regulates metabolic genes	81	3	1.41E-02	TXN, PRDX2, PRDX5
Detoxification of reactive oxygen species	32	3	1.84E-03	TXN, PRDX2, PRDX5

The cellular metabolism of amino acids covers mainly their synthesis and catabolism. The pathway “Selenoamino acid metabolism” presents a process in this area. In general, selenium represents a vital trace element in humans and it is for example incorporated into selenocysteine, where it replaces the sulphur atom. [107] The proteins adenosylhomocysteinase (*AHCY*), 40S ribosomal protein S3 (*RPS3*) and 60S acidic ribosomal protein P2 (*RPLP2*) are involved in the translational apparatus for the generation of selenoamino acids [108,109]. In detail, a study showed that adenosylhomocysteinase participates in the hydrolysis of adenosylselenohomocysteine into adenosine and selenohomocysteine [108]. Additionally, the ribosomal proteins are part of the selenocysteine translation machinery including the formation of selenocysteinyl-tRNA [109].

Further enriched pathways are correlated with the immune system, such as the “Innate immune system”, “Neutrophil degranulation”, as well as “Signaling by interleukins”. In detail, the proteins hemoglobin subunit beta (*HBB*), complement component 3 (*C3*) and galectin-3 (*LGALS3*) are involved in the process of neutrophil degranulation [110-112]. Neutrophils are a subtype of leukocytes and serve as first defensive line of the innate immune system [113]. In detail, the intrinsic granules of the neutrophils contain antimicrobial proteins, which are released via degranulation and function as potent response against intruders [114,115]. As an example, both complement component 3 and galectin-3 act as inflammatory mediators in the activation of neutrophils [111,112]. Another important molecule class for the immune system are interleukins, which are proteins that are involved in the intercellular communication between leukocytes [116]. As an example, the proteins cofilin-1, annexin A2, as well as vimentin interact with different types of interleukins and thus influence inflammatory responses [117-119].

The hit proteins galectin-3 (*LGALS3*), thioredoxin (*TXN*), peroxiredoxin-2 (*PRDX2*) and peroxiredoxin-5 (*PRDX5*) play a role in the pathway “RNA polymerase II transcription”, especially in the subcategory called “Generic transcription pathway”. In general, transcription is the process of gene expression via the synthesis of RNA from a DNA template and it can be divided into three major parts: initiation, elongation, and termination. Key players are the nuclear enzymes RNA polymerase I, II, and III. To elaborate, the RNA polymerase II transcribes primarily protein-coding genes and the activity as well as its regulation is crucial for the homeostasis of cells [120,121]. However, the enriched process “Generic transcription

pathway" includes mainly transcriptional regulation steps, in which transcription factors have a major impact as regulatory proteins [122]. For example, the transcription factor RUNX1 interacts with the promoter region of the LGALS3 gene and upregulates the transcription for the carbohydrate binding protein galectin-3 [123]. Generally, galectin-3 is included in several molecular processes such as cell adhesion and proliferation. Moreover, enhanced levels of galectin-3 have been associated with breast cancer [123,124].

Additional enriched pathways, which are included in the "Generic transcription pathway", are "Transcriptional regulation by *TP53*" and in particular "*TP53* regulates metabolic genes". The *TP53* gene encodes a protein called p53, which is also a transcription factor and functions as tumor suppressor. It controls cell division and cell growth by preventing uncontrolled cell proliferation and thus has an impact on the metabolism of carbohydrates, nucleotides, protein synthesis, as well as aerobic respiration [125]. As an example, p53 regulates the mitochondrial oxygen utilization and hence supports the reduction of molecular oxygen [126]. Furthermore, the proteins thioredoxin, peroxiredoxin-2 and peroxiredoxin-5 contribute to the redox environment of the cell by protecting it against oxidative stress, which can lead to DNA damage and genomic instability [127,128]. All three above mentioned proteins are also enriched in the pathways "Cellular responses to stress" and "Detoxification of reactive oxygen species". Especially noticeable is the fact that cancer cells often demonstrate an increased production of reactive oxygen species [128]. As a consequence, the redox cycle has been further investigated in many studies and differentially abundant levels of peroxiredoxin-2 and peroxiredoxin-5 have been detected in several cancers [128-130].

To sum up, in this chapter several molecular functions of the statistically significant proteins in the enriched pathway analysis are highlighted and potential correlations to cancer elaborated. In particular, the hit proteins are involved in inflammation processes, immune responses, transcription, translation, as well as maintenance of the cellular redox environment. Further evaluation and discussion of the results will be performed in chapter "10.2 Discussion".

10.1.2 Literature research

In the following chapter, a literature search of the hit proteins in the context of CRC is performed for additional interpretation. Of special interest is the determination of previously reported correlations between the proteins and CRC to evaluate the plausibility of the results and to examine possible identifications of proteins, which have no previous association to CRC. Here, a text mining R script called “OmixLitMiner”, which automatically executes a literature search based on PubMed Central® (https://www.ncbi.nlm.nih.gov/pubmed) for a given protein and keyword, is utilized [131]. In essence, the tool groups the proteins/genes into three main categories (1-3). Proteins/genes with category 1 are found in at least one review paper related to the searched context. For proteins/genes belonging to category 2 at least one publication is detected but no review paper. Category 3 is assigned to a protein/gene if no publication is discovered.

First, the literature mining was carried out with OmixLitMiner for the hit proteins/genes and the keywords “colorectal cancer” in a first search run and “colon cancer” in a second iteration. The assigned categories for a specific protein in each search run were compared and the smaller appointed category chosen for further interpretation. Next, a manual literature search was performed for the proteins for which no publication was found (category 3) in order to provide a potentially more thorough search. Here, the same categorization scheme was applied. To clarify, if a review paper was found for the target protein with the category 3 assignment, the category was altered to category 1 and if the manual search showed at least one publication, it was changed to category 2. An overview of the results is shown in table 8.

Table 8: Literature mining with the tool OmixLitMiner and manual curation. The first OmixLitMiner search (Search I) included the keyword "colorectal cancer" and the second search (Search II) the keyword "colon cancer". If the appointed categories differ, the smaller category is displayed under "final category". Manual literature mining was carried out for proteins, which had a category 3 assignment after the OmixLitMiner searches.

Hit Genes	Search "Colorectal cancer" Category	Search "Colon cancer" Category	Final Category
C3	2	2	2
S100P	2	2	2
PRDX2	2	2	2
ANXA1	2	3	2[a)]
VIM	2	3	2[a)]
TXN	2	3	2[a)]
LGALS3	2	3	2[a)]
RPS3	3	3	2[b)]
AHCY	3	3	2[b)]
CFL1	3	3	2[b)]
PRDX5	3	3	2[b)]
HBB	3	3	2[b)]
HBD	3	3	2[b)]
RPN1	3	3	2[b)]
RPLP2	3	3	2[b)]

a) Final category displays the smaller category assignment of the two OmixLitMiner searches.
b) Manually curated category assignments for previously appointed category 3 hits.

After the combination of an automated literature search via the OmixLitMiner tool and manual literature mining for the 15 statistically significant findings, table 8 demonstrates that all hit proteins have been previously described in the literature with a context to CRC.

10.2 Discussion

In the following chapter, the results of the biological inference regarding network and pathway analysis, as well as literature mining will be discussed, and potential merits and drawbacks examined.

As previously explained, only statistically significant findings which were identified across multiple search engines and stage-wise comparisons in each data analysis strategy were examined further. Hence, the biological inference was conducted by disregarding differences in the identification of statistically significant hits per stage-wise comparisons. Considering details provided by the stage-wise comparisons might be beneficial to further decipher different mechanisms in the development of CRC. In addition, data mining including other biological factors such as age, gender or location of the colorectal carcinoma might be interesting aspects to study. However, since DIA data mining demonstrated clear differences in the identification of statistically significant proteins per generated library, the authenticity of the findings was valued more than the potential benefit of considering detailed biological aspects. Consequently, only the hits verified by both data analysis strategies were used for biological inference.

In general, several tools for pathway and network analysis are available, such as STRING or DAVID [132,133]. Thereby, each tool uses different databases resulting into a distinction between the enriched pathways. Hence, each pathway and network analysis only serve as first indicator for establishing connections between the determined statistically significant proteins. However, the analysis via the reactome database shows several statistically enriched pathways (FDR < 5%). Crucial results include enrichment in inflammation processes, immune responses, transcription, translation, as well as maintenance of the cellular redox environment.

To further gain insights about the statistically significant proteins a literature search was performed based on a combination of the tool OmixLitMiner and manual literature mining. A particularly vital aspect is the identification of prior reported connections between the proteins and CRC to evaluate the plausibility of the results. Thereby, the automated search with the tool provides a useful overview about the findings. Since the mining with OmixLitMiner is dependent on the chosen keyword, such as “colorectal cancer”, it is beneficial to apply several iterations with related terms, such as “colon cancer”, in order to broaden the search space and thus to provide a more detailed output. In addition, further manual literature mining discovered a connection of all hit proteins to CRC. As a result, literature mining indicates a high degree of

plausibility for the identified significant proteins. However, while the combination of OmixLitMiner and manual literature mining provides a powerful approach to find meaningful results in a rather fast manner, it is important to note, that even for the combined strategy there is no guarantee of finding each possible connection between the hit proteins and individual search terms.

10.3 Conclusion

After the data analysis via the DDA-free and DDA-based strategy, 15 statistically significant proteins were analyzed regarding biological inference in order to achieve the main goal of the thesis - elucidate potential biological patterns in the context of CRC and highlight possible research targets for future studies.

The pathway and network analysis with the reactome database revealed many enriched biological paths (FDR < 5%) including inflammation processes, immune responses, as well as maintenance of the cellular redox environment. In addition, literature mining was performed to highlight potential research targets. After using the OmixLitMiner tool in combination with manual literature search, it was demonstrated that all the findings have a previous reported correlation to CRC, which adds plausibility to the results. After data analysis and biological inference all 15 hits display a high degree of authenticity. As a result, the here employed method including data analysis offers a promising strategy to detect a collection of significant proteins at once. Additionally, taking into account the vast heterogeneity of CRC, the identified protein panel is considered as a valuable starting point for a potential assay development to further unravel pathogenic mechanisms in CRC.

11. Concluding remarks & future perspectives

The main goal of the thesis was to identify significant proteins and elucidate potential biological patterns regarding CRC by comparing the protein profile of CRC samples in different stages. Hence, the thesis aimed at identifying a promising protein panel which can be used as a valuable starting point for further research. The method of choice for the detection of CRC-associated protein profiles was a label-free LC-MS/MS strategy with DIA. In addition, the primary focus was set on implementing a bioinformatic analysis workflow to exploit the high information input of the acquired digital DIA maps. The developed proteomic pipeline combined the results of multiple search engines to construct the corresponding libraries. Subsequently, the impact of each generated library on the DIA data analysis was investigated. Two different inputs were chosen for the bioinformatic workflow and the respective outcomes were compared: Pre-fractionated DDA measurements for the so called "DDA-based" analysis workflow and the DIA data for the termed "DDA-free" analysis strategy.

As a result, the thesis can be viewed from two main angles – bioinformatic analysis and biological inference.

Key element for both the DDA-based and DDA-free analysis is the importance of a specific-tailored spectral library for an efficient DIA data extraction. The resemblance of the library input and the DIA data is especially dependent on transition intensity and retention time variation. As stated previously, it might be beneficial to investigate post-measurement retention time alignment with CiRTs or DIAlignR and its impact on DIA extraction. In addition, the bioinformatic analysis revealed the direct influence of a chosen library on identified statistically significant proteins. As a conclusion, both the DDA-based and DDA-free strategy indicate that it can be beneficial to use the results of two database search engines and apply the whole analysis workflow for the search tools individually and their binary combination. Consequently, not only are the chances increased to extract valuable information provided by the DIA data, but the findings are also verified and the prospects to generate reproducible results are improved. In any case, the bioinformatic analysis stresses the importance of thorough deliberation in the interpretation of the acquired proteomics results.

The demonstrated differences in the DIA analysis impacted the analysis of the biological relevance. Only significant hits verified by both analysis strategies were directed to biological

inference, while disregarding potential valuable insights of the stage-wise comparisons. The authenticity of the statistically significant results was considered as most important. The findings are presented again in table 9 for a better overview. As mentioned previously, additional data mining including other biological factors such as gender, age, and location of the colorectal carcinoma might give further insights. However, the pathway and network analysis revealed enriched biological paths in inflammation processes, immune responses, and maintenance of the cellular redox environment. In addition, literature mining demonstrated that all identified proteins had a previously described association to CRC, which verified the findings and added plausibility. Hence, the applied method including the data analysis strategy provided the opportunity to discover a promising protein panel, which serves as a valuable starting point for a potential assay development in the ongoing research area of CRC. In this context, especially SRM-MS, which displays an excellent reproducibility and accuracy across large sample cohorts, is a valid method for further verifying the results and going the next step in the translation process from basic research to clinical utility [11].

Table 9: Protein panel as promising research targets in the context of CRC.

UniProtID	Protein names	Gene names
P23396	40S ribosomal protein S3	RPS3
P05387	60S acidic ribosomal protein P2	RPLP2
P23526	Adenosylhomocysteinase	AHCY
P04083	Annexin A1	ANXA1
P23528	Cofilin-1	CFL1
P01024	Complement C3	C3
P04843	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 1	RPN1
P17931	Galectin-3	LGALS3
P68871	Hemoglobin subunit beta	HBB
P02042	Hemoglobin subunit delta	HBD
P32119	Peroxiredoxin-2	PRDX2
P30044	Peroxiredoxin-5	PRDX5
P25815	Protein S100-P	S100P
P10599	Thioredoxin	TXN
P08670	Vimentin	VIM

As a conclusion, the applied DIA analysis with the developed proteomic pipeline is considered as a useful approach for discovering significant proteins and for providing the foundation for elucidating pathogenic patterns of CRC. In addition, the presented and evaluated proteomic pipeline offers further research studies an interesting method for balancing experimental against

computational costs. Especially in a clinical context, in which sample amounts, as well as measurement times represent limiting resources [3,134], the DDA-free analysis provides a promising strategy to generate meaningful results. Furthermore, it has been shown that the results in form of statistically significant proteins of the peptide-centric DIA data analysis are highly dependent on the used library, the applied filter settings during downstream analysis, and statistical considerations. Therefore, it can be beneficial to focus further research on the bioinformatic interplay of the different elements in DIA data analysis to achieve a higher level of reproducibility across proteomics results.

12. Materials and Methods

12.1 Instruments and Methods

All experimental procedures were conducted and all mass spectrometry raw data were provided by Niyusha Goudarzi Bozargomehri and Dr. Christoph Krisp. In total 70 human colorectal cancer (CRC) tissue samples were analyzed. In detail, 9 CRC samples were classified as stage I, 16 samples as stage II, 28 samples as stage III, and 17 samples as stage IV. From each tissue, ten 10 µm thick slices were prepared and transferred into an Eppendorf tube and subsequently stored at -80 °C.

12.1.1 Lysis, protein extraction, and in-solution proteolysis

The tissue samples were lysed in 200 µL of a SDC buffer (1% w/v sodium deoxycholate in 0.1 M triethylammonium bicarbonate) and sonicated for 10 seconds at 30% power. Subsequently, the samples were incubated at 98 °C for 5 min and the respective protein concentration was determined via a BCA Protein Assay (Pierce, Thermo Fisher Scientific). Next, SDC buffer was added to 20 µg of the lysate until a total volume of 100 µL was reached. In a following step, the disulfate bonds were reduced with 10 mM dithithreitol (DTT) at 60 °C for 30 min. Next, cycteine residues were alkylated in presence of 20 mM 2-iodoacetamide (IAA) at 37 °C for 30 min in the dark. The enzyme trypsin was added at a 1:50 ratio enzyme to protein and incubated at 37 °C overnight. The enzyme was inactivated and the SDC precipitated by the addition of 1 µL formic acid (FA) and a centrifugation was performed for 5 min at 14000 g. The supernatant was collected and dried in a vacuum concentrator.

12.1.2 HpH-reversed phase chromatography for spectral library generation

Five microgram of each digested CRC samples of stages I and II, as well as the samples of stage III and stage IV were taken and combined together and the pH adjusted to 10.5 with a ammonium hydroxide solution. Next, peptides were separated within a 25 min gradient (3 – 35% acetonitrile) with a flow rate of 200 µL/min on a monolith column (ProSwift™ RP-4H, 1 mm x 250 mm, Thermo Fisher Scientific) using an HPLC system (Agilent 1200 series, Agilent Technologies) with a two buffer system. Equilibration buffer: 5 mM ammonium

hydroxide (pH = 10.5), elution buffer 5 mM ammonium hydroxide in 90% acetonitrile. A fraction collector was used to collect 1 min fractions beginning from the first minute until 29 fractions were collected. The collected fractions were combined according to the following scheme - fractions 1-3, 4-6, 7-9, 10+20, 11+21, 12+22, 13+23, 14+24, 15+25, 16+26, 17+27, 18+28 and 19+29. As a result, 13 fractions were acquired per pool. In total, 26 samples were dried and then dissolved in 25 µL (0.1%) FA for MS analysis.

12.1.3 LC method for DDA and DIA experiments

For both DDA and DIA analysis the same LC set-up and parameters were used. Per MS measurement 1 µg of protein sample was injected. The flowrate was set to 0.25 µL/min. The gradient started with 98% buffer A (99.9% H20 and 0.1% FA) and 2% buffer B (100% ACN and 0.1% FA). The concentration of buffer B was increased to 30% within 60 min and subsequently enhanced to 95% within 1 min. The concentration stayed constant for 5 min and was decreased to 2% within 1 min. The concentration was held constant for 15 min before measuring the next sample.

12.1.4 MS parameter for the DDA experiments

Samples were analyzed on a Quadrupole Orbitrap hybrid mass spectrometer (QExactive, Thermo Fisher Scientific). The data were acquired in data-dependent mode. The analysis time was 75 min. Fullscan spectra were acquired in profile mode utilizing a resolution of 70000 with a scan range of 400 to 1300 m/z and an accumulation time of 120 ms. Fragment spectra were acquired applying a resolution of 17500 with an accumulation time of 60 ms. The dynamic exclusion for precursor ions was set to 20 ms. Ions were fragmented using higher energy collisional dissociation (HCD) with stepped normalized collision energy (25 and 28). Signals with a single charge or more than 5 charges were excluded from fragmentation.

12.1.5 MS parameter for the DIA experiments

Samples were analyzed on a Quadrupole Orbitrap hybrid mass spectrometer (QExactive, Thermo Fisher Scientific). The data were acquired in data-independent mode. The analysis time was 75 min. Fullscan spectra were acquired in a scan range of 390 to 1210 m/z with a resolution of 70000 and an accumulation time of 55 ms. For MS2 acquisition a resolution of 17500 was used and a window size of 25 m/z was applied from 400 to 1200 m/z. Fragment ion spectra were accumulated for 100 ms and acquired with stepped normalized collision energy (27, 28, and 29). A detailed list of the applied DIA windows is shown in table 10.

Table 10: DIA windows with start and end values.

Start m/z	End m/z	Start m/z	End m/z
400.44	425.44	800.62	825.62
425.45	450.45	825.63	850.63
450.46	475.46	850.64	875.64
475.47	500.47	875.65	900.65
500.48	525.48	900.67	925.67
525.49	550.49	925.68	950.68
550.51	575.51	950.69	975.69
575.52	600.52	975.70	1000.7
600.53	625.53	1000.71	1025.71
625.54	650.54	1025.72	1050.72
650.55	675.55	1050.73	1075.73
675.56	700.56	1075.74	1100.74
700.57	725.57	1100.76	1125.76
725.59	750.59	1125.77	1150.77
750.60	775.60	1150.78	1175.78
775.61	800.61	1175.79	1200.79

12.2 Data Analysis

12.2.1 Peptide identification, library generation, and DIA analysis

26 DDA raw-files and 70 DIA raw-files were used for the data analysis. DDA and DIA data were converted and centroided with MSConvert (version 3.0.9134). The DIA data was further processed with DIA-Umpire (version 2.0) to generate pseudo-MS/MS spectra, which were submitted to the database search. Peptide identification was performed with the database search engines Comet (version 2016.01 rev. 3), MS-GF+ (version 2018.01.30), and X!Tandem (version 2015.12.15) and a reviewed human FASTA database retrieved from UniProtKB in November 2018. Decoy entries were generated via the module DecoyDatabase (version 2.3.0) and appended to the FASTA file. Key search parameters included trypsin as enzyme and one allowed missed cleavage, a precursor mass tolerance of 10 ppm, as well as precursor charges from 2 to 5. Cystein carbamidomethylation was applied as fixed modification and oxidation on methionine was set as variable modification. Further validation of the assigned peptide-spectrum matches was performed with PeptideProphet (version 5.1.0), iProphet (version 5.1.0) and MAYU (version 1.07). The design of an automated process up to library generation was accomplished via batch scripts. Library generation and the subsequent DIA analysis was applied with Skyline (version 19.1) including filter criteria such as a protein FDR < 1%, precursor charges from 2 to 4, ion charges from 1 to 3, as well as only using y- and b-ions. High-quality DIA data extraction was based on a dotP > 0.8 and at least 2 peptides per protein.

12.2.2 Statistical analysis, network analysis, and literature mining

For statistical analysis, the data was processed with MSstats (version 3.18.4). Normalization was performed based on equalized medians and the calculation of summed up Log-intensities relied on a Tukey-Median Polish. Further statistical analysis was applied in the R environment (version 3.6.0) including a Welch’s t-test and a subsequent Benjamin-Hochberg multiple hypothesis testing correction. Further data mining, as well as visualization was performed with self-written R scripts. Pathway and network analysis were applied with ReactomeFI in the cytoscape environment (version 3.7.0). Literature mining included the OmixLitMiner tool.

13. References

[1] C. Coghlin, G. Murray. Biomarkers of colorectal cancer: recent advances and future challenges. **2015**, *Proteomics. Clinical applications*, 1-2, 64–71.

[2] M. Gonzalez-Pons, M. Cruz-Correa. Colorectal Cancer Biomarkers: Where Are We Now? **2015**, *BioMed research international*, 149014.

[3] H. Ma, G. Chen, M. Guo. Mass spectrometry based translational proteomics for biomarker discovery and application in colorectal cancer. **2016**, *Proteomics. Clinical applications*, 4, 503–515.

[4] P. Aghagolzadeh, R. Radpour. New trends in molecular and cellular biomarker discovery for colorectal cancer. **2016**, *World journal of gastroenterology*, 25, 5678–5693.

[5] R. Nibbe, M. Chance. Approaches to biomarkers in human colorectal cancer: looking back, to go forward. **2009**, *Biomarkers in medicine*, 4, 385–396.

[6] J. Obuch, D. Ahnen. Colorectal Cancer: Genetics is Changing Everything. **2016**, *Gastroenterology clinics of North America*, 3, 459–476.

[7] V. Das, J. Kalita, M. Pal. Predictive and prognostic biomarkers in colorectal cancer: A systematic review of recent advances and challenges. **2017**, *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, 8–19.

[8] Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. **2001**, *Clinical pharmacology and therapeutics*, 3, 89–95.

[9] N. Goossens, *et al.* Cancer biomarker discovery and validation. **2015**, *Translational cancer research*, 3, 256–269.

[10] C. Jimenez, *et al.* Proteomics of colorectal cancer: overview of discovery studies and identification of commonly identified cancer-associated proteins and candidate CRC serum markers. **2010**, *Journal of proteomics*, 10, 1873–1895.

[11] M. de Wit, *et al.* Proteomics in colorectal cancer translational research: biomarker discovery for clinical applications. **2013**, *Clinical biochemistry*, 6, 466–479.

[12] G. Lech, *et al.* Colorectal cancer tumour markers and biomarkers: Recent therapeutic advances. **2016**, *World journal of gastroenterology*, 5, 1745–1755.

[13] A. Alnabulsi, G. Murray. Proteomics for early detection of colorectal cancer: recent updates. **2018**, *Expert review of proteomics*, 1, 55–63.

[14] R. Aebersold. A stress test for mass spectrometry-based proteomics. **2009**, *Nature methods*, 6, 411–412.

[15] R. Aebersold, M. Mann. Mass spectrometry-based proteomics. **2003**, *Nature*, 6928, 198–207.

[16] L. Smith, N. Kelleher. Proteoform: a single term describing protein complexity. **2013**, *Nature methods*, 3, 186–187.

[17] T. Toby, L. Fornelli, N. Kelleher. Progress in Top-Down Proteomics and the Analysis of Proteoforms. **2016**, *Annual review of analytical chemistry (Palo Alto, Calif.)*, 1, 499–519.

[18] J. Tran, *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. **2011**, *Nature*, 7376, 254–258.

[19] R. Aebersold, M. Mann. Mass-spectrometric exploration of proteome structure and function. **2016**, *Nature*, 7620, 347–355.

[20] Y. Liu, *et al.* Mass spectrometric protein maps for biomarker discovery and clinical research. **2013**, *Expert review of molecular diagnostics*, 8, 811–825.

[21] D. Tabb, *et al.* Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. **2010**, *Journal of proteome research*, 2, 761–776.

[22] T. Shi, *et al.* Advances in targeted proteomics and applications to biomedical research. **2016**, *Proteomics*, 15-16, 2160–2182.

[23] E. Borràs, E. Sabidó. What is targeted proteomics? A concise revision of targeted acquisition and targeted data analysis in mass spectrometry. **2017**, *Proteomics*, 17-18.

[24] R. Bruderer, *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. **2015**, *Molecular & cellular proteomics : MCP*, 5, 1400–1410.

[25] K. Blackburn, *et al.* Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation. **2010**, *Journal of proteome research*, 7, 3621–3637.

[26] J. Muntel, *et al.* Advancing Urinary Protein Biomarker Discovery by Data-Independent Acquisition on a Quadrupole-Orbitrap Mass Spectrometer. **2015**, *Journal of proteome research*, 11, 4752–4762.

[27] L. Gillet, *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. **2012**, *Molecular & cellular proteomics : MCP*, 6, O111.016717.

[28] C. Ludwig, *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. **2018**, *Molecular systems biology*, 8, e8126.

[29] C. Colangelo, *et al.* Review of software tools for design and analysis of large scale MRM proteomic datasets. **2013**, *Methods (San Diego, Calif.)*, 3, 287–298.

[30] A. Keller, *et al.* A uniform proteomics MS/MS analysis platform utilizing open XML file formats. **2005**, *Molecular systems biology*, 2005.0017.

[31] J. Cox, M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. **2008**, *Nature Biotechnology*, 12, 1367–1372.

[32] C. Masselon, *et al.* Accurate mass multiplexed tandem mass spectrometry for high-throughput polypeptide identification from mixtures. **2000**, *Analytical chemistry*, 8, 1918–1924.

[33] J. Venable, *et al.* Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. **2004**, *Nature methods*, 1, 39–45.

[34] A. Bilbao, *et al.* Processing strategies and software solutions for data-independent acquisition in mass spectrometry. **2015**, *Proteomics*, 5-6, 964–980.

[35] N. Selevsek, *et al.* Reproducible and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-mass spectrometry. **2015**, *Molecular & cellular proteomics : MCP*, 3, 739–749.

[36] J. Zi, *et al.* Expansion of the ion library for mining SWATH-MS data through fractionation proteomics. **2014**, *Analytical chemistry*, 15, 7242–7246.

[37] E. Govaert, *et al.* Comparison of fractionation proteomics for local SWATH library building. **2017**, *Proteomics*, 15-16.

[38] M. Matsumoto, *et al.* A large-scale targeted proteomics assay resource based on an in vitro human proteome. **2017**, *Nature methods*, 3, 251–258.

[39] G. Rosenberger, *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. **2014**, *Scientific data*, 140031.

[40] O. Schubert, *et al.* Absolute Proteome Composition and Dynamics during Dormancy and Resuscitation of Mycobacterium tuberculosis. **2015**, *Cell host & microbe*, 1, 96–108.

[41] S. Anjo, C. Santa, B. Manadas. SWATH-MS as a tool for biomarker discovery: From basic research to clinical applications. **2017**, *Proteomics*, 3-4.

[42] Y. Luo, *et al.* SWATH-based proteomics identified carbonic anhydrase 2 as a potential diagnosis biomarker for nasopharyngeal carcinoma. **2017**, *Scientific reports*, 41191.

[43] O. Schubert, *et al.* Building high-quality assay libraries for targeted analysis of SWATH MS data. **2015**, *Nature protocols*, 3, 426–441.

[44] Y. Ting, *et al.* Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. **2015**, *Molecular & cellular proteomics : MCP*, 9, 2301–2307.

[45] L. Gillet, A. Leitner, R. Aebersold. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. **2016**, *Annual review of analytical chemistry (Palo Alto, Calif.)*, 1, 449–472.

[46] W. Noble, M. MacCoss. Computational and statistical analysis of protein mass spectrometry data. **2012**, *PLoS computational biology*, 1, e1002296.

[47] A. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. **2010**, *Journal of proteomics*, 11, 2092–2123.

[48] C. Hughes, B. Ma, G. Lajoie. De novo sequencing methods in proteomics. **2010**, *Methods in molecular biology (Clifton, N.J.)*, 105–121.

[49] J. Taylor, R. Johnson. Sequence database searches viade novo peptide sequencing by tandem mass spectrometry. **1997**, *Rapid Communications in Mass Spectrometry*, 9, 1067–1075.

[50] A. Frank, P. Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. **2005**, *Analytical chemistry*, 4, 964–973.

[51] K. Vyatkina, *et al.* Top-down analysis of protein samples by de novo sequencing techniques. **2016**, *Bioinformatics (Oxford, England)*, 18, 2753–2759.

[52] T. Muth, B. Renard. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? **2018**, *Briefings in bioinformatics*, 5, 954–970.

[53] H. Lam, *et al.* Building consensus spectral libraries for peptide identification in proteomics. **2008**, *Nature methods*, 10, 873–875.

[54] J. Griss. Spectral library searching in proteomics. **2016**, *Proteomics*, 5, 729–740.

[55] B. Frewen, *et al.* Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. **2006**, *Analytical chemistry*, 16, 5678–5684.

[56] M. Gentzel, *et al.* Preprocessing of tandem mass spectrometric data to support automatic protein identification. **2003**, *Proteomics*, 8, 1597–1610.

[57] H. Xu, M. Freitas. A dynamic noise level algorithm for spectral screening of peptide MS/MS spectra. **2010**, *BMC bioinformatics*, 436.

[58] R. Sadygov, D. Cociorva, J. Yates. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. **2004**, *Nature methods*, 3, 195–202.

[59] K. Verheggen, *et al.* Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. **2017**, *Mass spectrometry reviews*.

[60] J. Eng, *et al.* A fast SEQUEST cross correlation algorithm. **2008**, *Journal of proteome research*, 10, 4598–4602.

[61] J. Eng, T. Jahan, M. Hoopmann. Comet: an open-source MS/MS sequence database search tool. **2013**, *Proteomics*, 1, 22–24.

[62] R. Craig, R. Beavis. TANDEM: matching proteins with tandem mass spectra. **2004**, *Bioinformatics (Oxford, England)*, 9, 1466–1467.

[63] S. Dasari, *et al.* TagRecon: high-throughput mutation identification through sequence tagging. **2010**, *Journal of proteome research*, 4, 1716–1726.

[64] S. Kim, P. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. **2014**, *Nature communications*, 5277.

[65] V. Bafna, N. Edwards. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. **2001**, *Bioinformatics (Oxford, England)*, S13-21.

[66] D. Shteynberg, *et al.* Combining results of multiple search engines in proteomics. **2013**, *Molecular & cellular proteomics : MCP*, 9, 2383–2393.

[67] J. Paulo. Practical and Efficient Searching in Proteomics: A Cross Engine Comparison. **2013**, *WebmedCentral*, 10.

[68] J. Cottrell. Protein identification using MS/MS data. **2011**, *Journal of proteomics*, 10, 1842–1851.

[69] M. The, *et al.* Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. **2016**, *Journal of the American Society for Mass Spectrometry*, 11, 1719–1727.

[70] K. Ma, O. Vitek, A. Nesvizhskii. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. **2012**, *BMC bioinformatics*, S1.

[71] A. Kertesz-Farkas, *et al.* Database Searching in Mass Spectrometry Based Proteomics. **2012**, *Current Bioinformatics*, 2, 221–230.

[72] D. Shteynberg, *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. **2011**, *Molecular & cellular proteomics : MCP*, 12, M111.007690.

[73] L. Reiter, *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. **2009**, *Molecular & cellular proteomics : MCP*, 11, 2405–2417.

[74] E. Audain, *et al.* In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. **2017**, *Journal of proteomics*, 170–182.

[75] O. Serang, M. MacCoss, W. Noble. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. **2010**, *Journal of proteome research*, 10, 5346–5357.

[76] S. Sikdar, R. Gill, S. Datta. Improving protein identification from tandem mass spectrometry data by one-step methods and integrating data from other platforms. **2016**, *Briefings in bioinformatics*, 2, 262–269.

[77] B. MacLean, *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. **2010**, *Bioinformatics (Oxford, England)*, 7, 966–968.

[78] H. Röst, *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. **2014**, *Nature Biotechnology*, 219 EP -.

[79] J. Meyer, B. Schilling. Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques. **2017**, *Expert review of proteomics*, 5, 419–429.

[80] C.-C. Tsou, *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. **2015**, *Nature methods*, 3, 258-64, 7 p following 264.

[81] Y. Li, *et al.* Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. **2015**, *Nature methods*, 12, 1105–1106.

[82] R. Bruderer, *et al.* High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. **2016**, *Proteomics*, 15-16, 2246–2256.

[83] H. Röst, *et al.* TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. **2016**, *Nature methods*, 9, 777–783.

[84] L. Reiter, *et al.* mProphet: automated data processing and statistical validation for large-scale SRM experiments. **2011**, *Nature methods*, 5, 430–435.

[85] J. Teleman, *et al.* DIANA--algorithmic improvements for analysis of data-independent acquisition MS data. **2015**, *Bioinformatics (Oxford, England)*, 4, 555–562.

[86] U. Toprak, *et al.* Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. **2014**, *Molecular & cellular proteomics : MCP*, 8, 2056–2071.

[87] M. Choi, *et al.* MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. **2014**, *Bioinformatics (Oxford, England)*, 17, 2524–2526.

[88] M. Narasimhan, *et al.* Clinical biomarker discovery by SWATH-MS based label-free quantitative proteomics: impact of criteria for identification of differentiators and data normalization method. **2019**, *Journal of translational medicine*, 1, 184.

[89] S. Tyanova, *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. **2016**, *Nature methods*, 9, 731–740.

[90] P. Navarro, *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. **2016**, *Nature Biotechnology*, 11, 1130–1136.

[91] P. Abramowski, *et al.* Combined Application of RGB Marking and Mass Spectrometruc Imaging Facilitates Detection of Tumor Heterogeneity. **2015**, *Cancer genomics & proteomics*, 12, 179–187.

[92] M. Aichler, A. Walch. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. **2015**, *Laboratory Investigation*, 4, 422–431.

[93] A. Hinsch, *et al.* MALDI imaging mass spectrometry reveals multiple clinically relevant masses in colorectal cancer using large-scale tissue microarrays. **2017**, *Journal of Mass Spectrometry*, 3, 165–173.

[94] S. Parker, *et al.* Identification of a Set of Conserved Eukaryotic Internal Retention Time Standards for Data-independent Acquisition Mass Spectrometry. **2015**, *Molecular & cellular proteomics : MCP*, 10, 2800–2813.

[95] S. Gupta, *et al.* DIAlignR provides precise retention time alignment across distant runs in DIA and targeted proteomics. **2018**,

[96] A. Fabregat, *et al.* The Reactome Pathway Knowledgebase. **2018**, *Nucleic acids research*, D1, D649-D655.

[97] B. Jassal, *et al.* The reactome pathway knowledgebase. **2020**, *Nucleic acids research*, D1, D498-D503.

[98] G. Wu, R. Haw. Functional Interaction Network Construction and Analysis for Disease Discovery. **2017**, *Methods in molecular biology (Clifton, N.J.)*, 235–253.

[99] R. Gungabissoon, J. Bamburg. Regulation of growth cone actin dynamics by ADF/cofilin. **2003**, *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, 4, 411–420.

[100] M. Koppers, *et al.* Receptor-specific interactome as a hub for rapid cue-induced selective translation in axons. **2019**, *eLife*.

[101] G. Lolas, A. Bianchi, K. Syrigos. Tumour-induced neoneurogenesis and perineural tumour growth: a mathematical approach. **2016**, *Scientific reports*, 20684.

[102] D. Palm, F. Entschladen. Neoneurogenesis and the neuro-neoplastic synapse. **2007**, *Progress in experimental tumor research*, 91–98.

[103] Y. Nyathi, B. Wilkinson, M. Pool. Co-translational targeting and translocation of proteins to the endoplasmic reticulum. **2013**, *Biochimica et biophysica acta*, 11, 2392–2402.

[104] D. Akopian, *et al.* Signal recognition particle: an essential protein-targeting machine. **2013**, *Annual review of biochemistry*, 693–721.

[105] S. Jagannathan, *et al.* Multifunctional roles for the protein translocation machinery in RNA anchoring to the endoplasmic reticulum. **2014**, *The Journal of biological chemistry*, 37, 25907–25924.

[106] K. Pogue-Geile, *et al.* Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein. **1991**, *Molecular and cellular biology*, 8, 3842–3849.

[107] R. Schmidt, M. Simonović. Synthesis and decoding of selenocysteine and human health. **2012**, *Croatian medical journal*, 6, 535–550.

[108] E. Kajander, *et al.* Metabolism, cellular actions, and cytotoxicity of selenomethionine in cultured cells. **1991**, *Biological trace element research*, 1, 57–68.

[109] P. O'Donoghue, J. Ling, D. Söll. Transfer RNA function and evolution. **2018**, *RNA biology*, 4-5, 423–426.

[110] G. Lominadze, *et al.* Proteomic analysis of human neutrophil granules. **2005**, *Molecular & cellular proteomics : MCP*, 10, 1503–1521.

[111] G. Fernández, *et al.* Galectin-3 and soluble fibrinogen act in concert to modulate neutrophil activation and survival: involvement of alternative MAPK pathways. **2005**, *Glycobiology*, 5, 519–527.

[112] L. Camous, *et al.* Complement alternative pathway acts as a positive feedback amplification of neutrophil activation. **2011**, *Blood*, 4, 1340–1349.

[113] L. Felix, S. Almas, P. Lacy. Regulatory Mechanisms in Neutrophil Degranulation. **2018**, *Immunopharmacology and Inflammation*, 191–210.

[114] C. Yin, B. Heit. Armed for destruction: formation, function and trafficking of neutrophil granules. **2018**, *Cell and tissue research*, 3, 455–471.

[115] V. Rungelrath, S. Kobayashi, F. DeLeo. Neutrophils in innate immunity and systems biology-level approaches. **2020**, *Wiley interdisciplinary reviews. Systems biology and medicine*, 1, e1458.

[116] M. Akdis, *et al.* Interleukins (from IL-1 to IL-38), interferons, transforming growth factor β, and TNF-α: Receptors, functions, and roles in diseases. **2016**, *The Journal of allergy and clinical immunology*, 4, 984–1010.

[117] A. Hirayama, *et al.* Cofilin plays a critical role in IL-8-dependent chemotaxis of neutrophilic HL-60 cells through changes in phosphorylation. **2007**, *Journal of leukocyte biology*, 3, 720–728.

[118] E. Solito, *et al.* IL-6 stimulates annexin 1 expression and translocation and suggests a new biological role as class II acute phase protein. **1998**, *Cytokine*, 7, 514–521.

[119] P. Hornbeck, *et al.* Vimentin expression is differentially regulated by IL-2 and IL-4 in murine T cells. **1993**, *Journal of immunology (Baltimore, Md. : 1950)*, 8, 4013–4021.

[120] N. Fuda, M. Ardehali, J. Lis. Defining mechanisms that regulate RNA polymerase II transcription in vivo. **2009**, *Nature*, 7261, 186–192.

[121] V. Svetlov, E. Nudler. Basic mechanism of transcription by RNA polymerase II. **2013**, *Biochimica et biophysica acta*, 1, 20–28.

[122] S. Sperling. Transcriptional regulation at a glance. **2007**, *BMC bioinformatics*, S2.

[123] H.-Y. Zhang, *et al.* RUNX1 and RUNX2 upregulate Galectin-3 expression in human pituitary tumors. **2009**, *Endocrine*, 1, 101–111.

[124] H. Zhang, *et al.* Galectin-3 as a marker and potential therapeutic target in breast cancer. **2014**, *PloS one*, 9, e103482.

[125] K. Sullivan, *et al.* Mechanisms of transcriptional regulation by p53. **2018**, *Cell death and differentiation*, 1, 133–143.

[126] J. Fields, *et al.* How does p53 regulate mitochondrial respiration? **2007**, *IUBMB life*, 10, 682–684.

[127] Y. Kim, H. Jang. Role of Cytosolic 2-Cys Prx1 and Prx2 in Redox Signaling. **2019**, *Antioxidants (Basel, Switzerland)*, 6.

[128] A. Nicolussi, *et al.* The role of peroxiredoxins in cancer. **2017**, *Molecular and clinical oncology*, 2, 139–153.

[129] C. Cadenas, *et al.* Role of thioredoxin reductase 1 and thioredoxin interacting protein in prognosis of breast cancer. **2010**, *Breast cancer research : BCR*, 3, R44.

[130] W. Lu, *et al.* Peroxiredoxin 2 is upregulated in colorectal cancer and contributes to colorectal cancer cells' survival by protecting cells from oxidative stress. **2014**, *Molecular and cellular biochemistry*, 1-2, 261–270.

[131] P. Steffen, *et al.* OmixLitMiner: A Bioinformatics Tool for Prioritizing Biological Leads from 'Omics Data Using Literature Retrieval and Data Mining. **2020**, *International journal of molecular sciences*, 4.

[132] D. Huang, *et al.* DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. **2007**, *Nucleic acids research*, Web Server issue, W169-75.

[133] D. Szklarczyk, *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. **2019**, *Nucleic acids research*, D1, D607-D613.

[134] J. Forshed. Experimental Design in Clinical 'Omics Biomarker Discovery. **2017**, *Journal of proteome research*, 11, 3954–3960.

14. Appendix

14.1 GHS classification of the chemicals

No chemicals were used.

14.2 DDA-based analysis - Volcano plots of stage-wise comparisons

For each stage-wise comparison a Welch's t-test and a Benjamin-Hochberg multiple hypothesis testing correction were performed (FDR < 5%). The results are displayed via volcano plots (Fig. 56 - Fig. 79). Statistically significant proteins are marked in red.

14.2.1 Stage I vs. Stage II

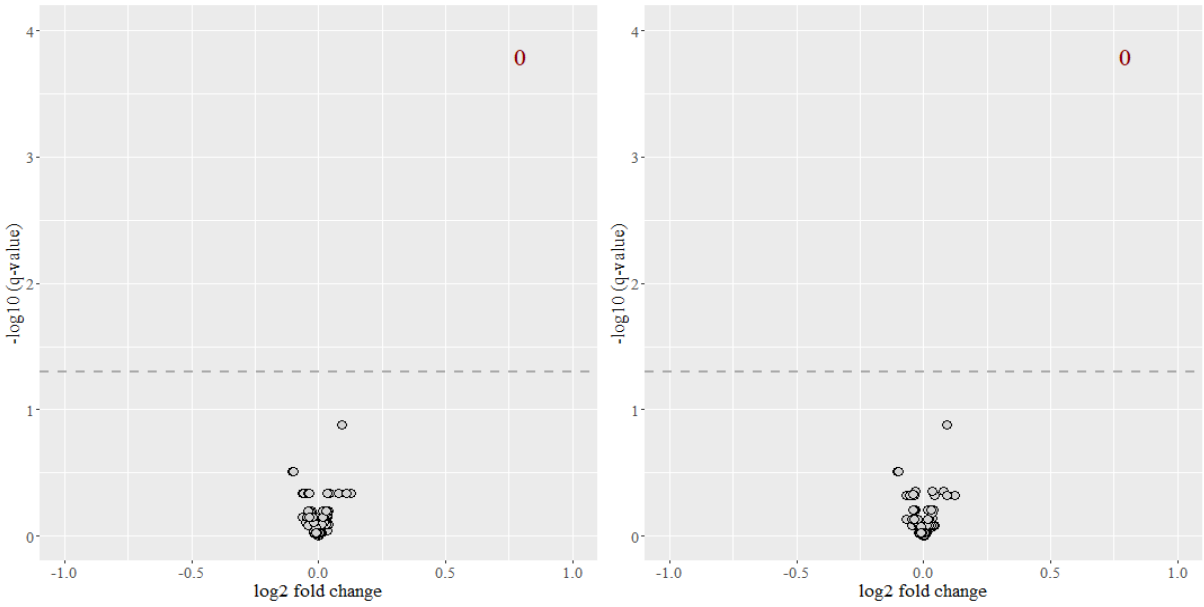


Fig. 56: Volcano plots for the comparison of Stage I and Stage II for C (left) and M (right).

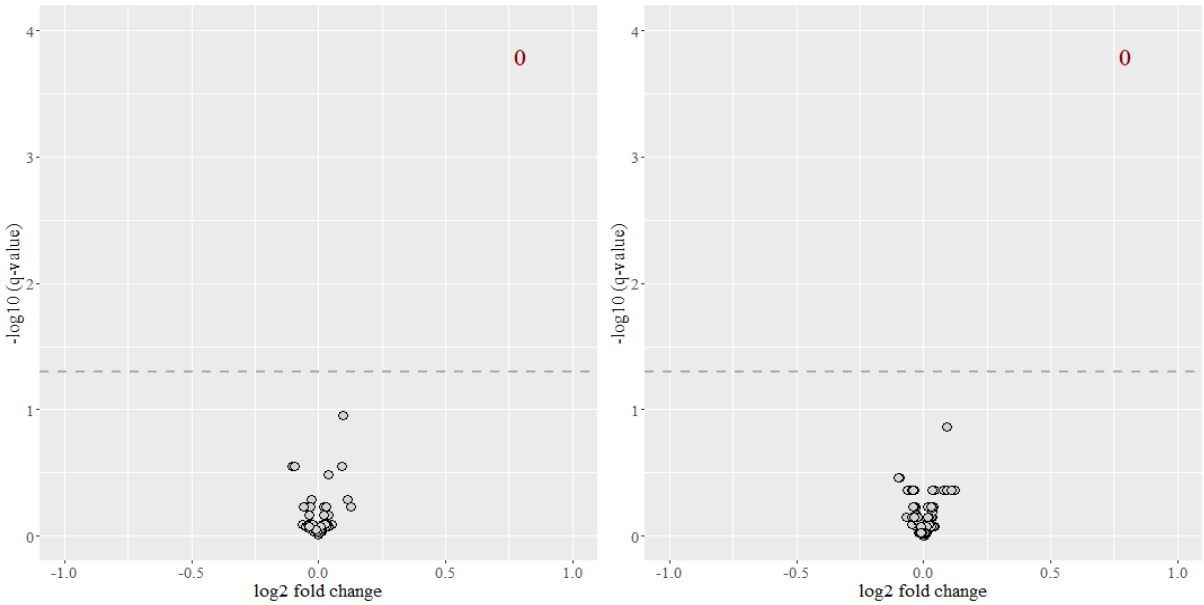


Fig. 57: Volcano plots for the comparison of Stage I and Stage II for T (left) and the combination of CM (right).

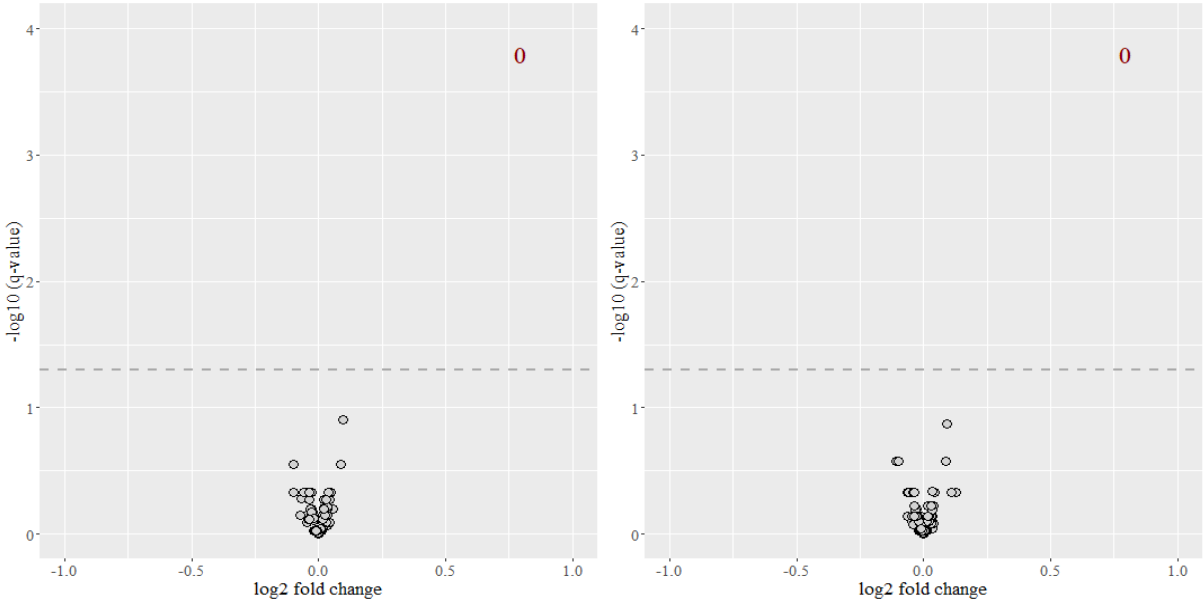


Fig. 58: Volcano plots for the comparison of Stage I and Stage II for the combination of CT (left) and the combination of MT (right).

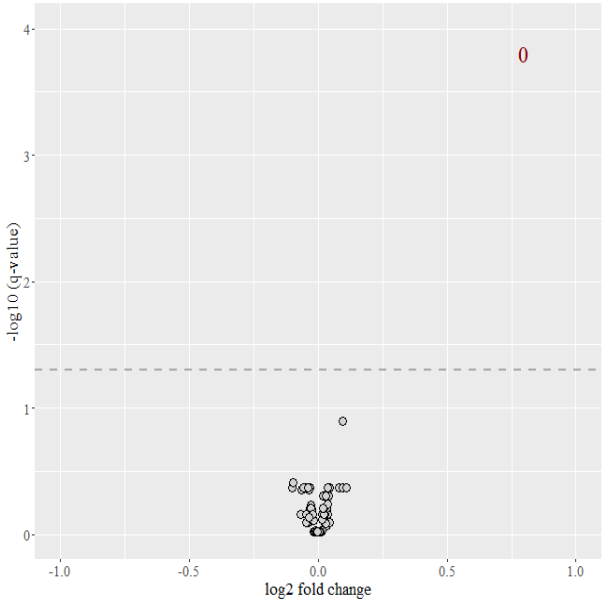


Fig. 59: Volcano plot for the comparison of Stage I and Stage II for the combination of CMT .

14.2.2 Stage I vs. Stage III

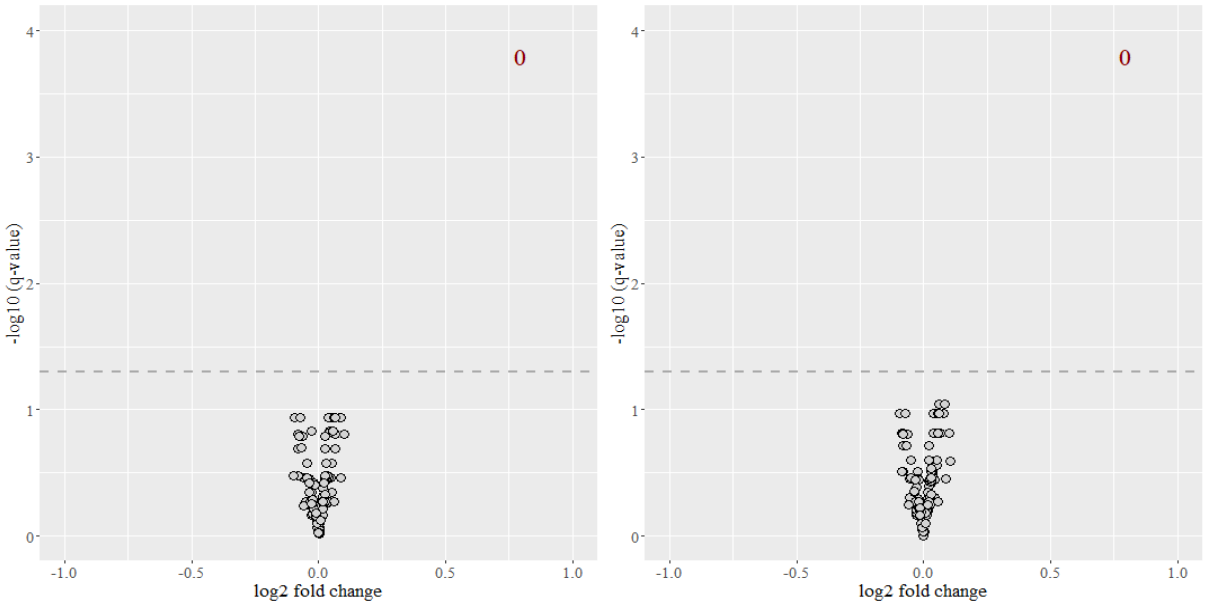


Fig. 60: Volcano plots for the comparison of Stage I and Stage III for C (left) and M (right).

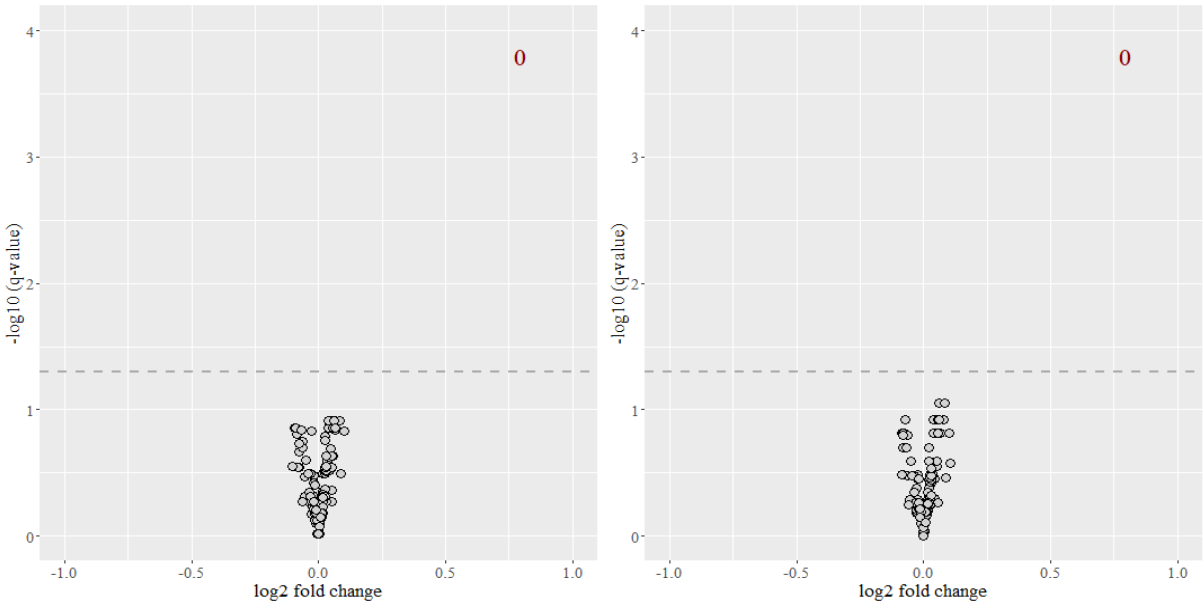


Fig. 61: Volcano plots for the comparison of Stage I and Stage III for T (left) and the combination of CM (right).

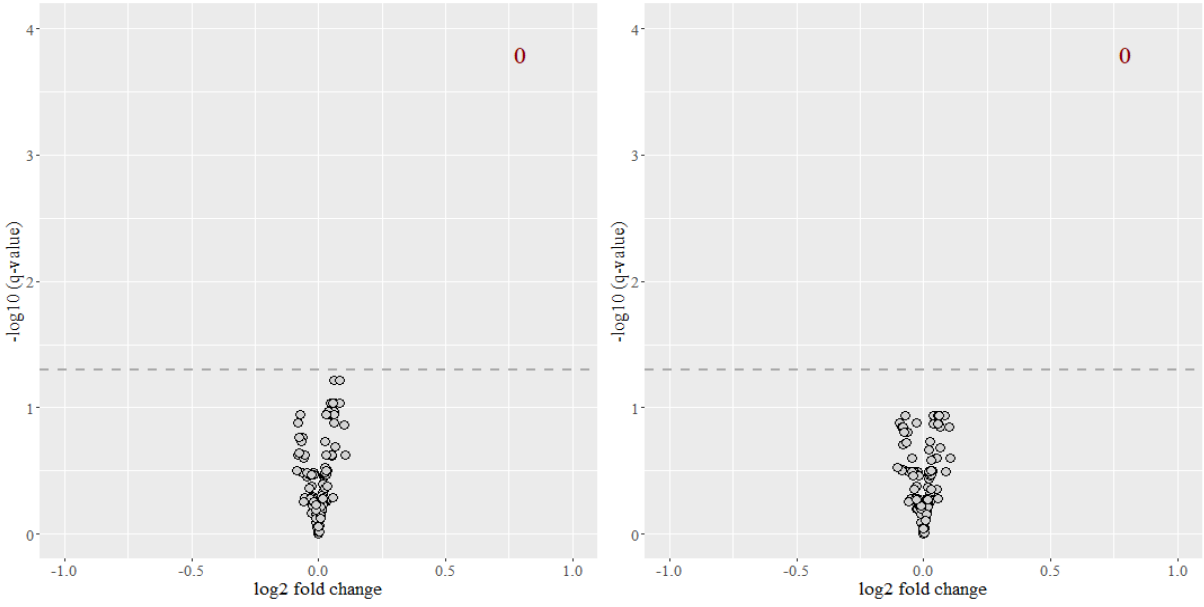


Fig. 62: Volcano plots for the comparison of Stage I and Stage III for the combination of CT (left) and the combination of MT (right).

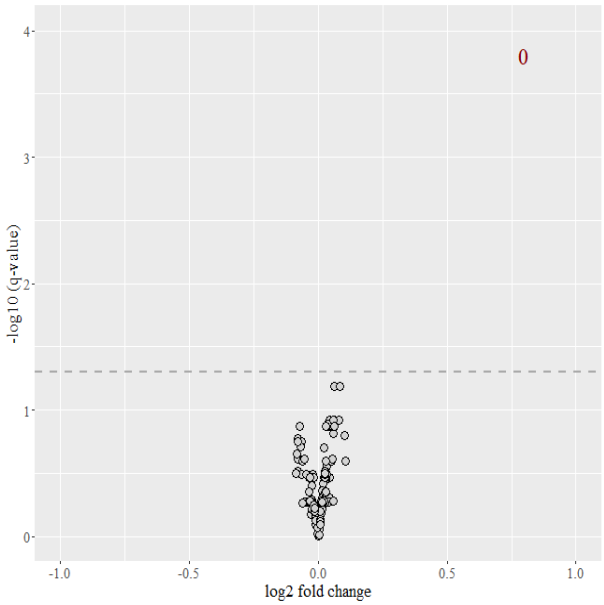


Fig. 63: Volcano plot for the comparison of Stage I and Stage III for the combination of CMT .

14.2.3 Stage I vs. Stage IV

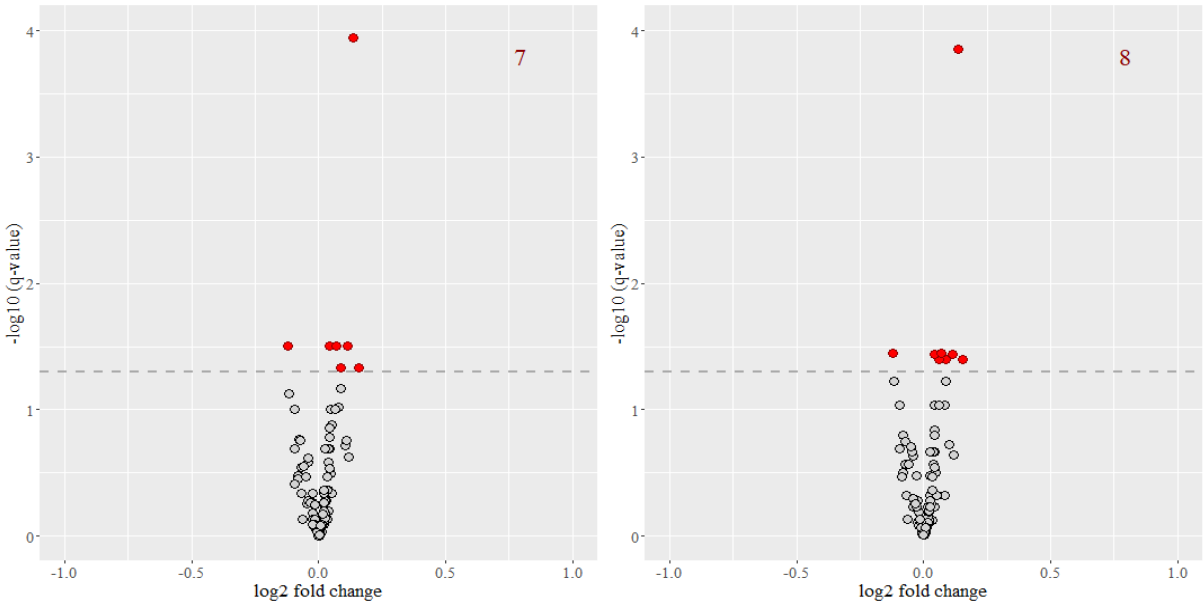


Fig. 64: Volcano plots for the comparison of Stage I and Stage IV for C (left) and M (right).

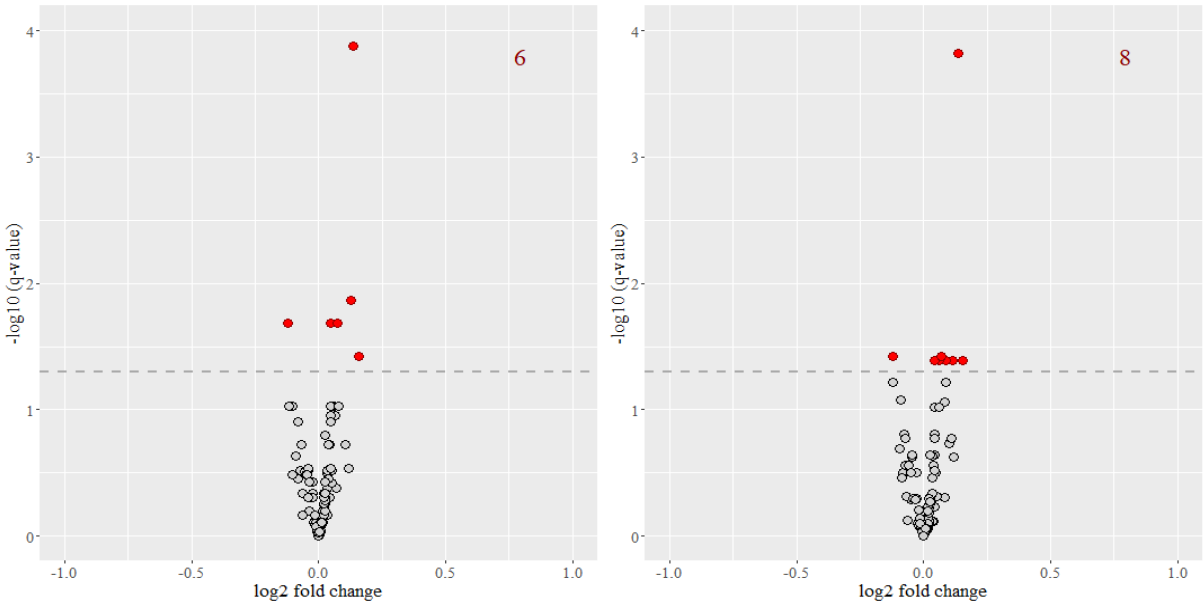


Fig. 65: Volcano plots for the comparison of Stage I and Stage IV for T (left) and the combination of CM (right).

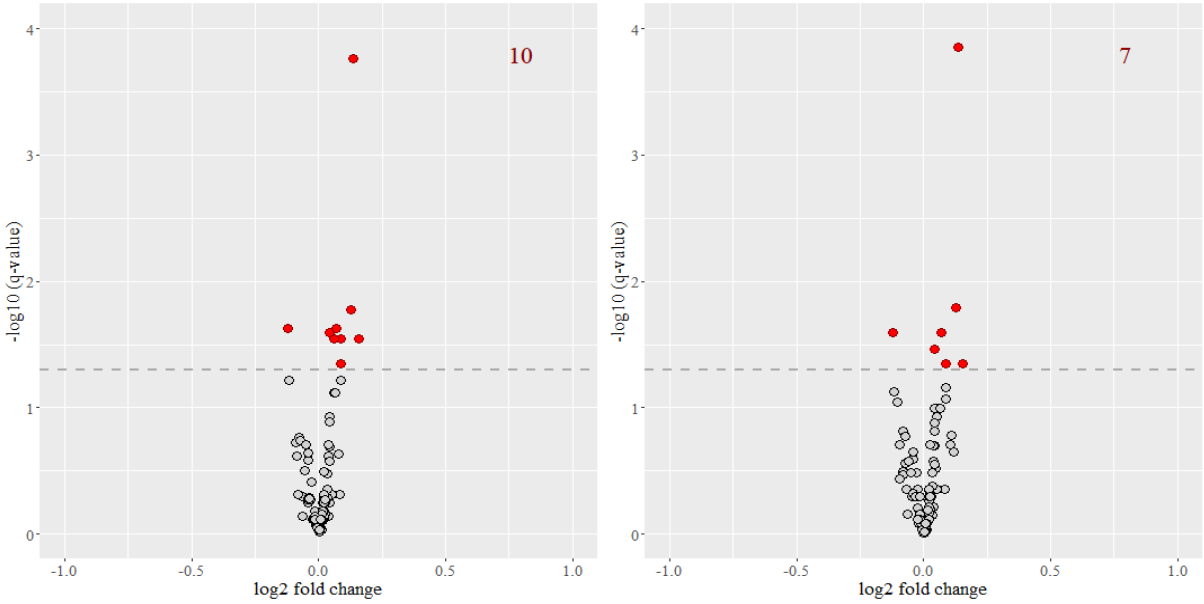


Fig. 66: Volcano plots for the comparison of Stage I and Stage IV for the combination of CT (left) and the combination of MT (right).

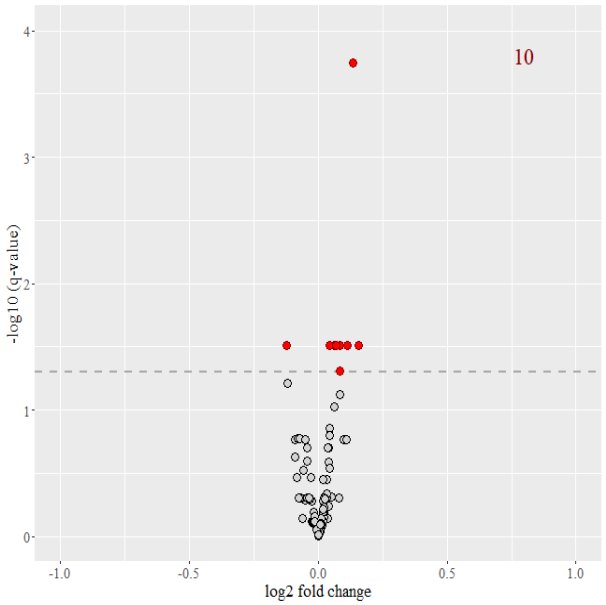


Fig. 67: Volcano plot for the comparison of Stage I and Stage IV for the combination of CMT .

14.2.4 Stage II vs. Stage III

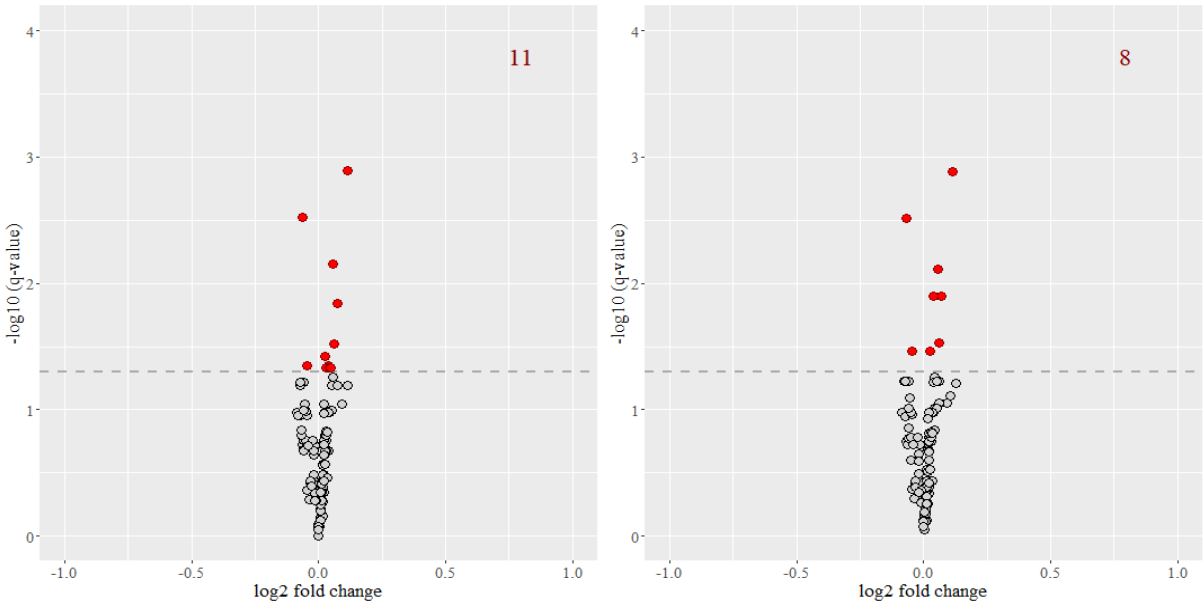


Fig. 68: Volcano plots for the comparison of Stage II and Stage III for C (left) and M (right).

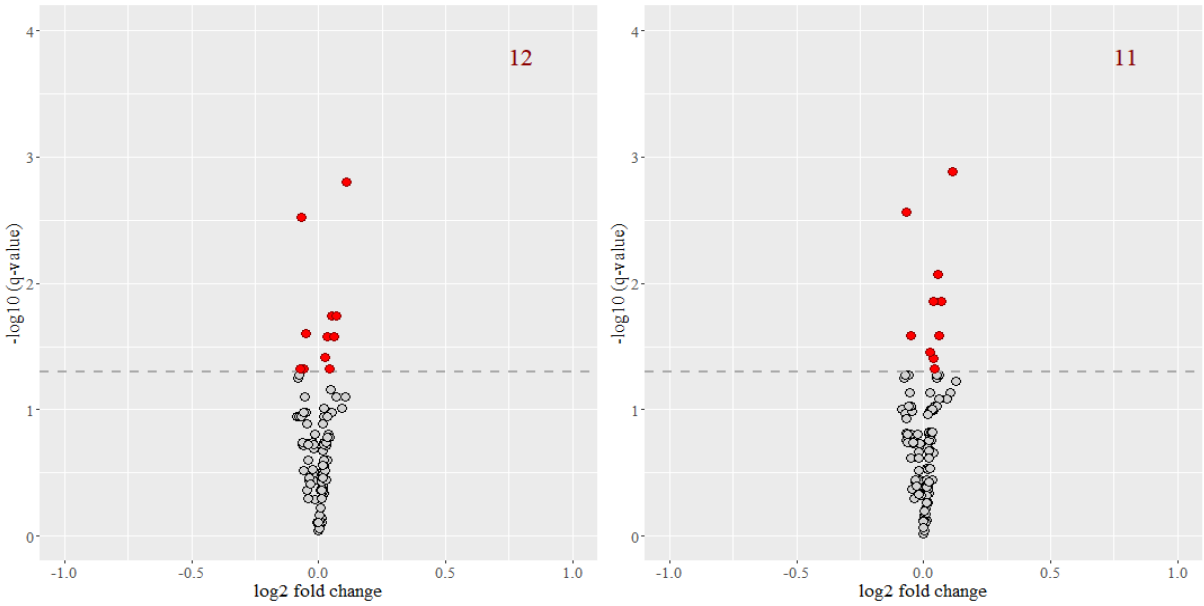


Fig. 69: Volcano plots for the comparison of Stage II and Stage III for T (left) and the combination of CM (right).

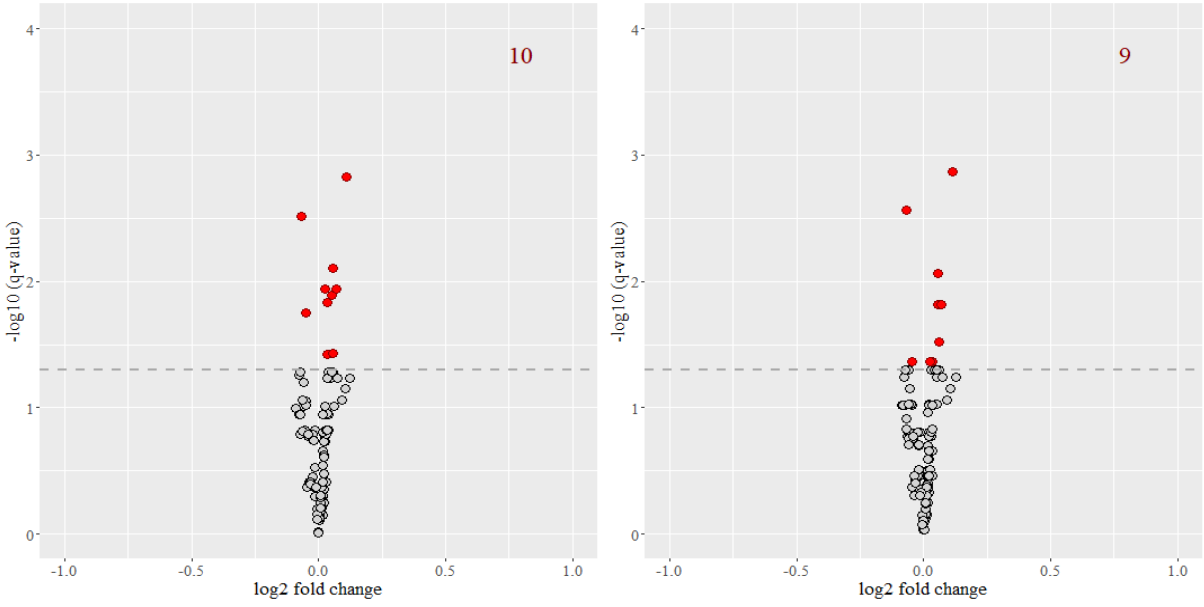


Fig. 70: Volcano plots for the comparison of Stage II and Stage III for the combination of CT (left) and the combination of MT (right).

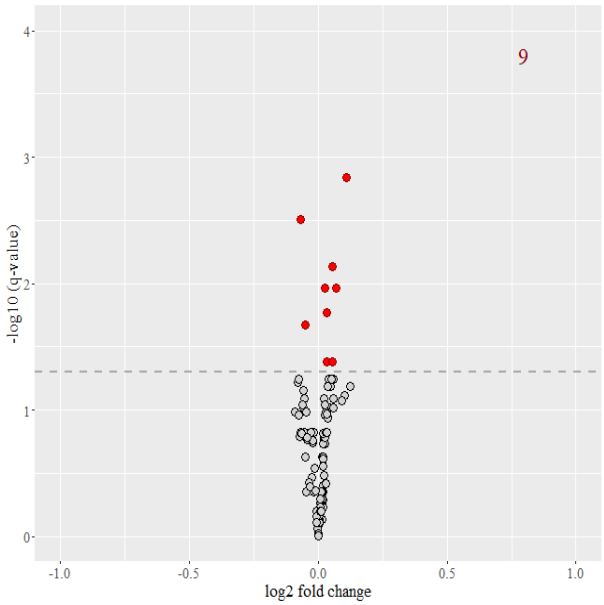


Fig. 71: Volcano plot for the comparison of Stage II and Stage III for the combination of CMT .

14.2.5 Stage II vs. Stage IV

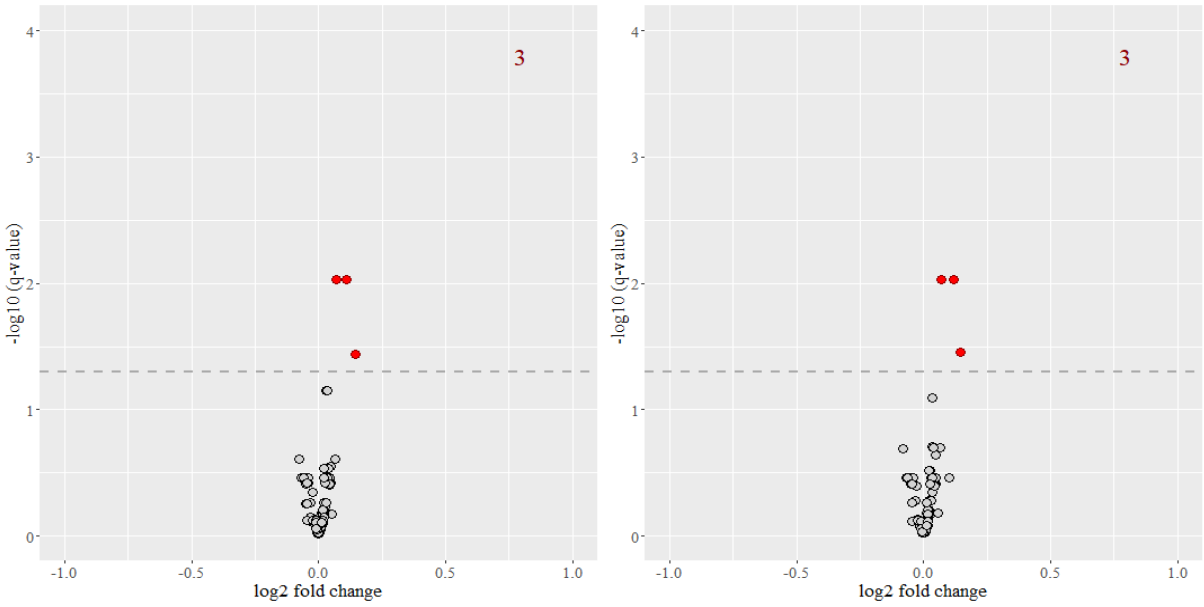


Fig. 72: Volcano plots for the comparison of Stage II and Stage IV for C (left) and M (right).

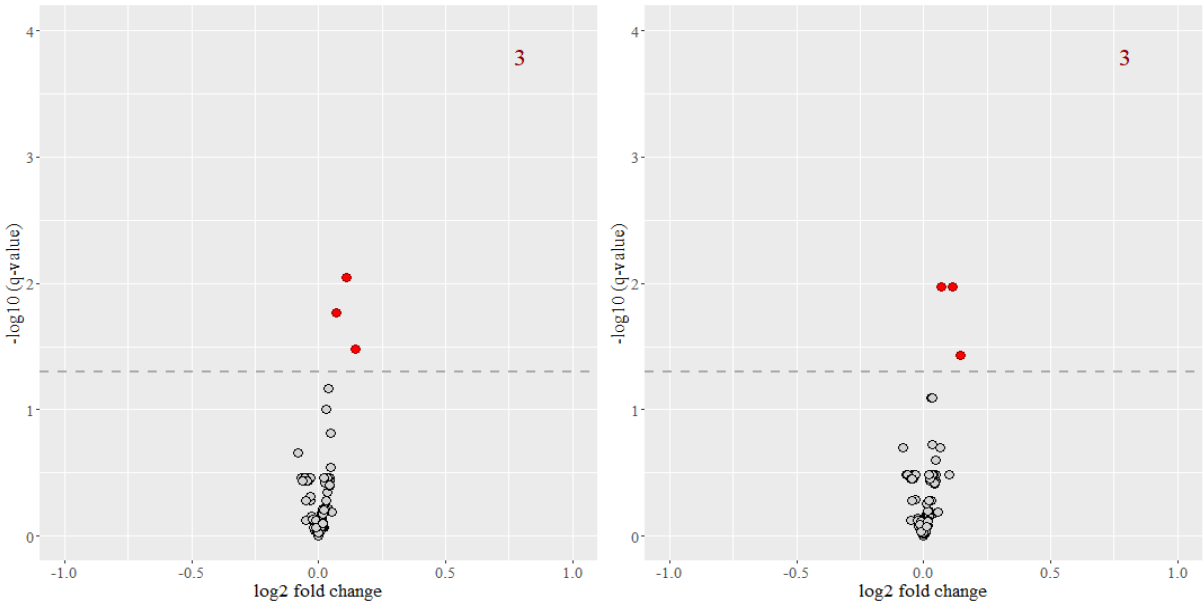


Fig. 73: Volcano plots for the comparison of Stage II and Stage IV for T (left) and the combination of CM (right).

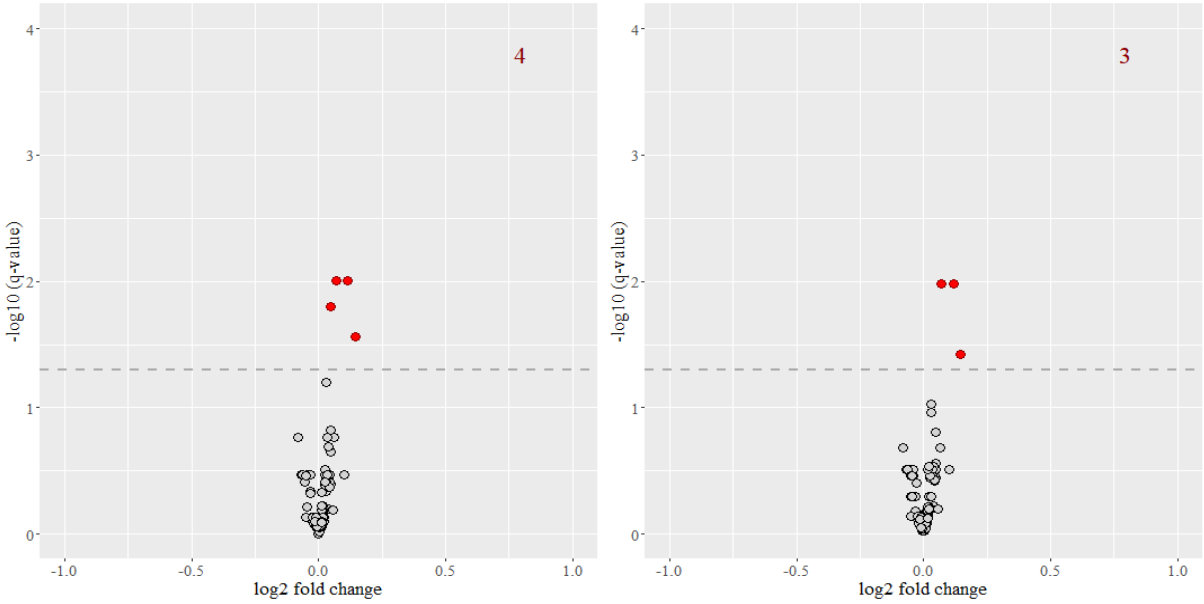


Fig. 74: Volcano plots for the comparison of Stage II and Stage IV for the combination of CT (left) and the combination of MT (right).

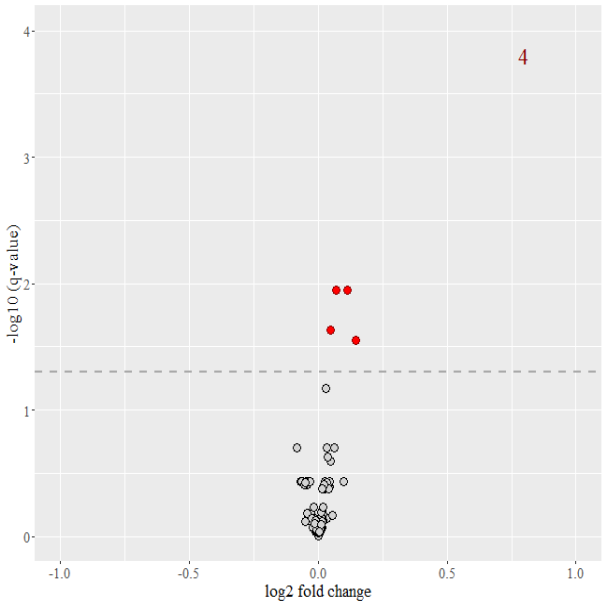


Fig. 75: Volcano plot for the comparison of Stage II and Stage IV for the combination of CMT .

14.2.6 Stage III vs. Stage IV

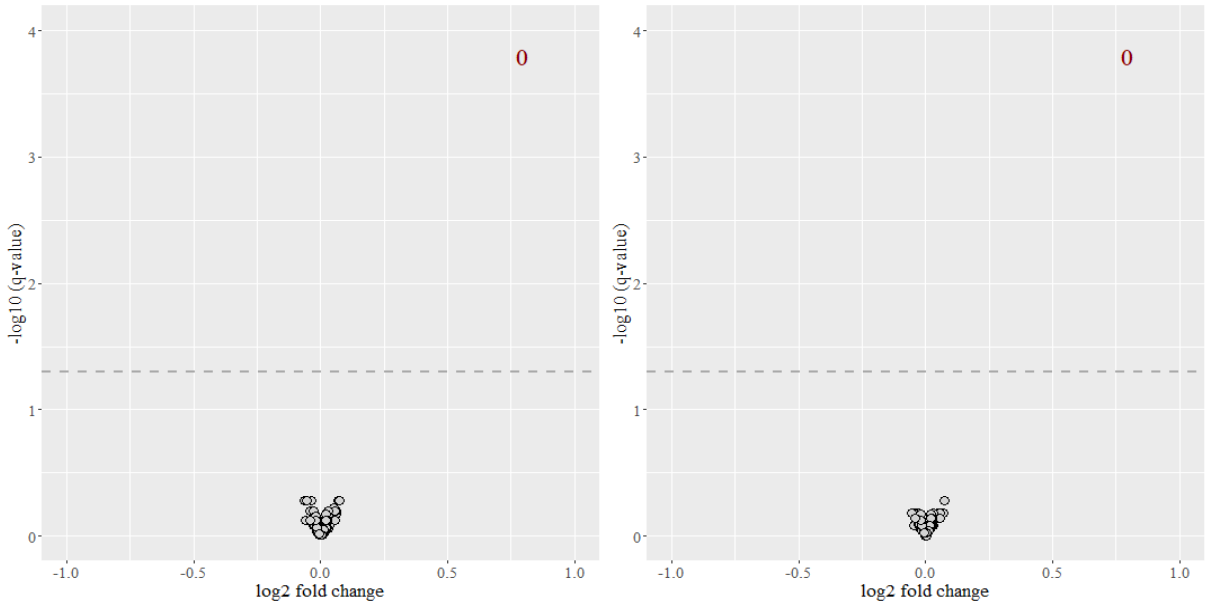


Fig. 76: Volcano plots for the comparison of Stage III and Stage IV for C (left) and M (right).

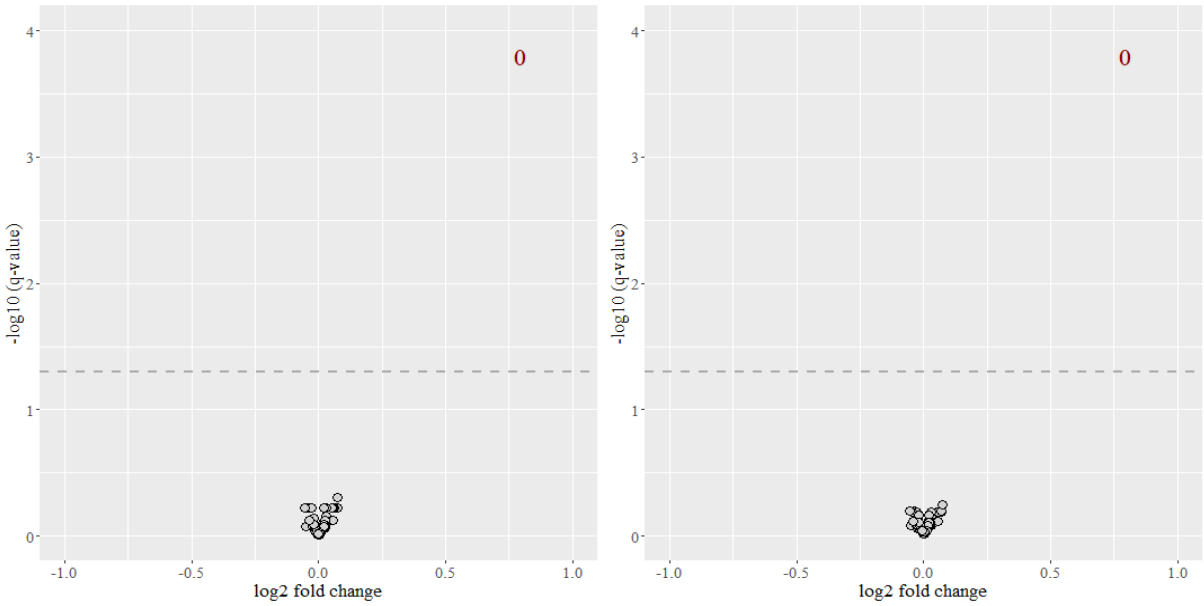


Fig. 77: Volcano plots for the comparison of Stage III and Stage IV for T (left) and the combination of CM (right).

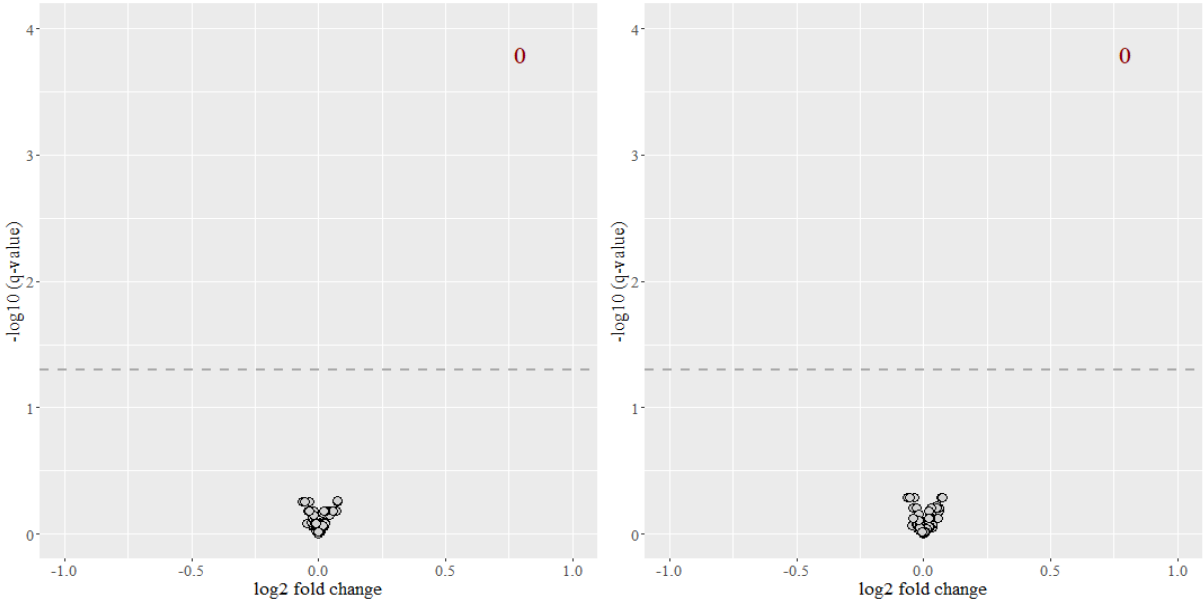


Fig. 78: Volcano plots for the comparison of Stage III and Stage IV for the combination of CT (left) and the combination of MT (right).

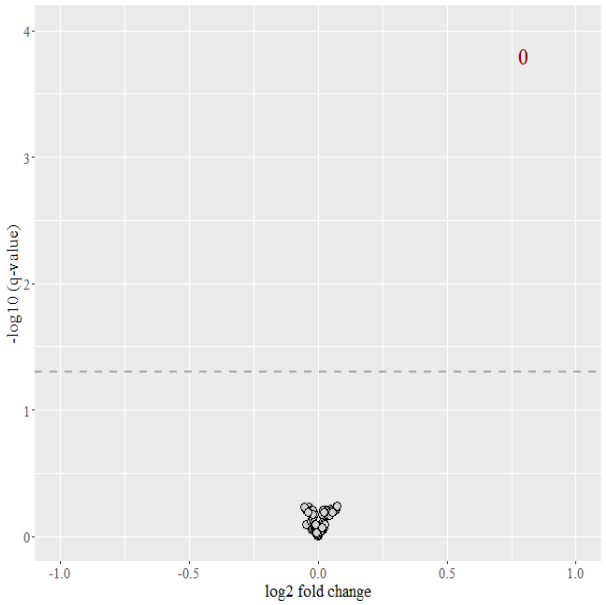


Fig. 79: Volcano plot for the comparison of Stage III and Stage IV for the combination of CMT .

14.3 DDA-free analysis - Volcano plots of stage-wise comparisons

For each stage-wise comparison a Welch's t-test and a Benjamin-Hochberg multiple hypothesis testing correction were performed (FDR < 5%). The results are displayed via volcano plots (Fig. 80 - Fig. 103). Statistically significant proteins are marked in red.

14.3.1 Stage I vs. Stage II

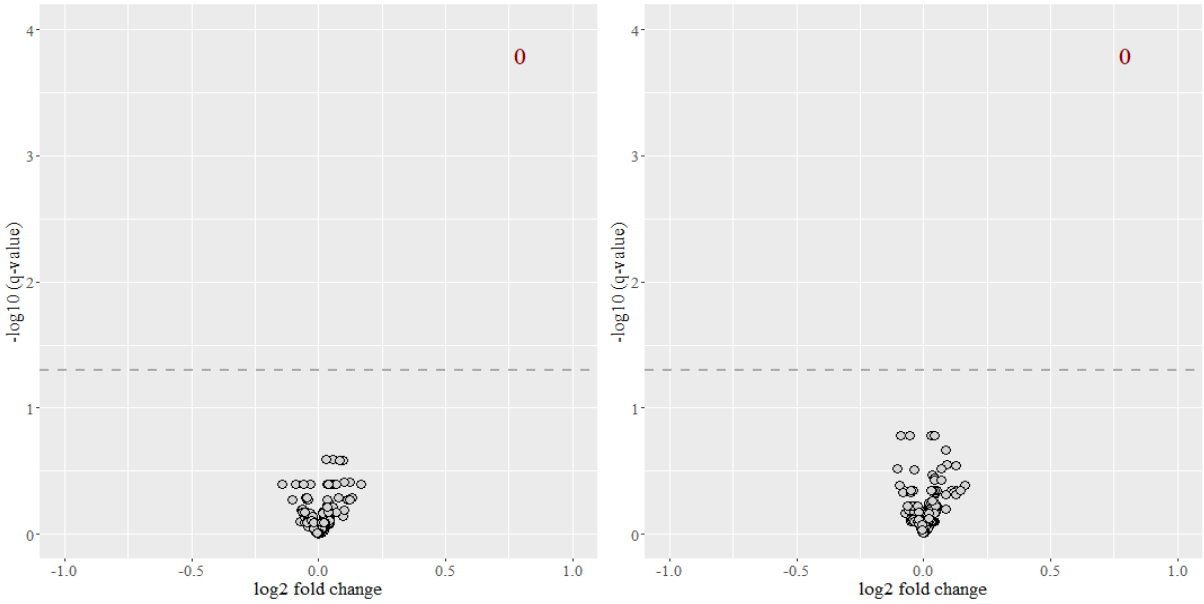


Fig. 80: Volcano plots for the comparison of Stage I and Stage II for C (left) and M (right).

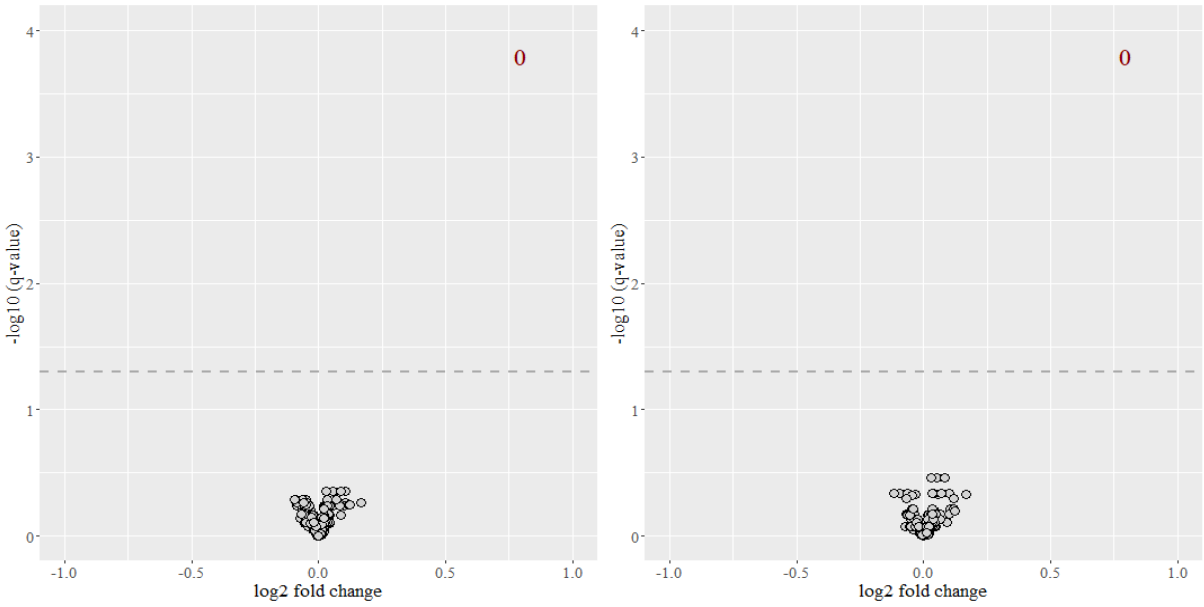


Fig. 81: Volcano plots for the comparison of Stage I and Stage II for T (left) and the combination of CM (right).

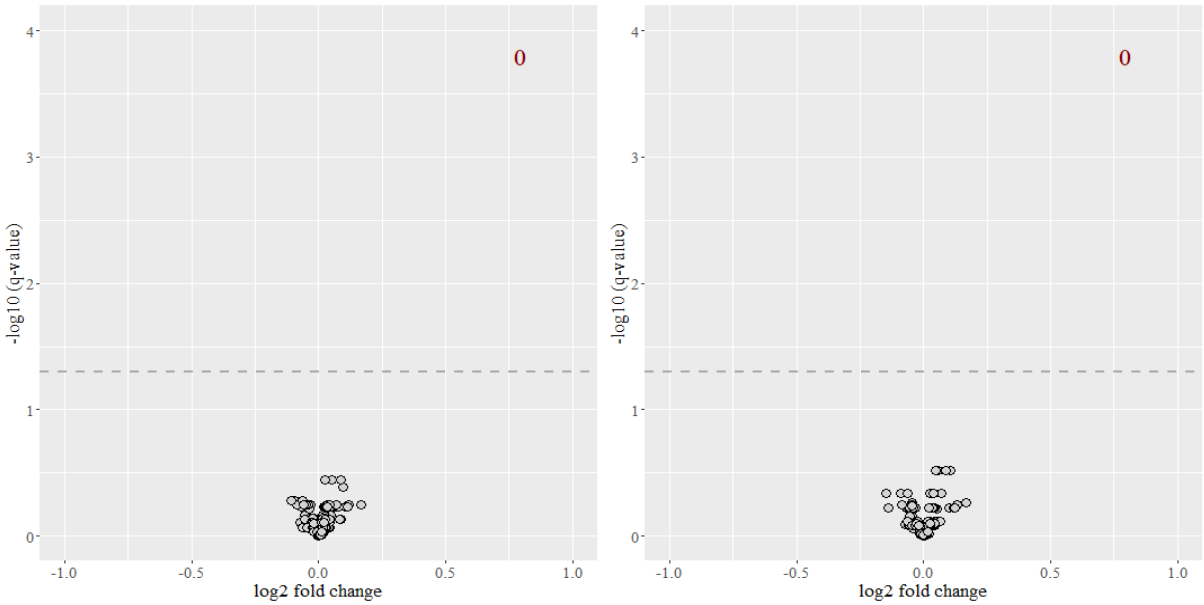


Fig. 82: Volcano plots for the comparison of Stage I and Stage II for the combination of CT (left) and the combination of MT (right).

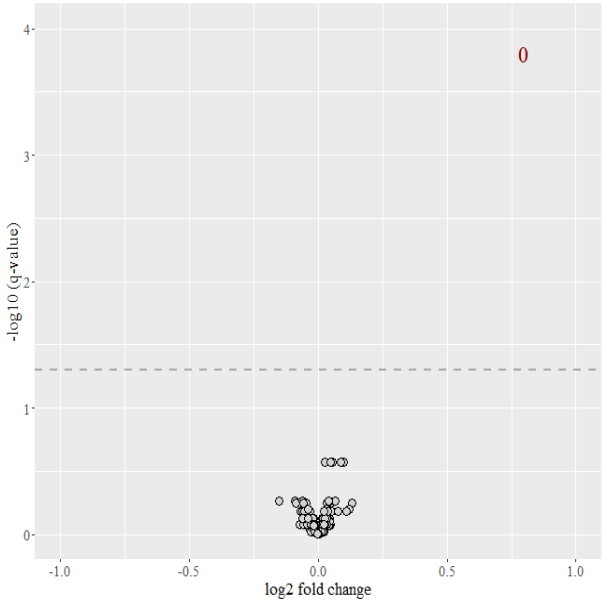


Fig. 83: Volcano plot for the comparison of Stage I and Stage II for the combination of CMT .

14.3.2 Stage I vs. Stage III

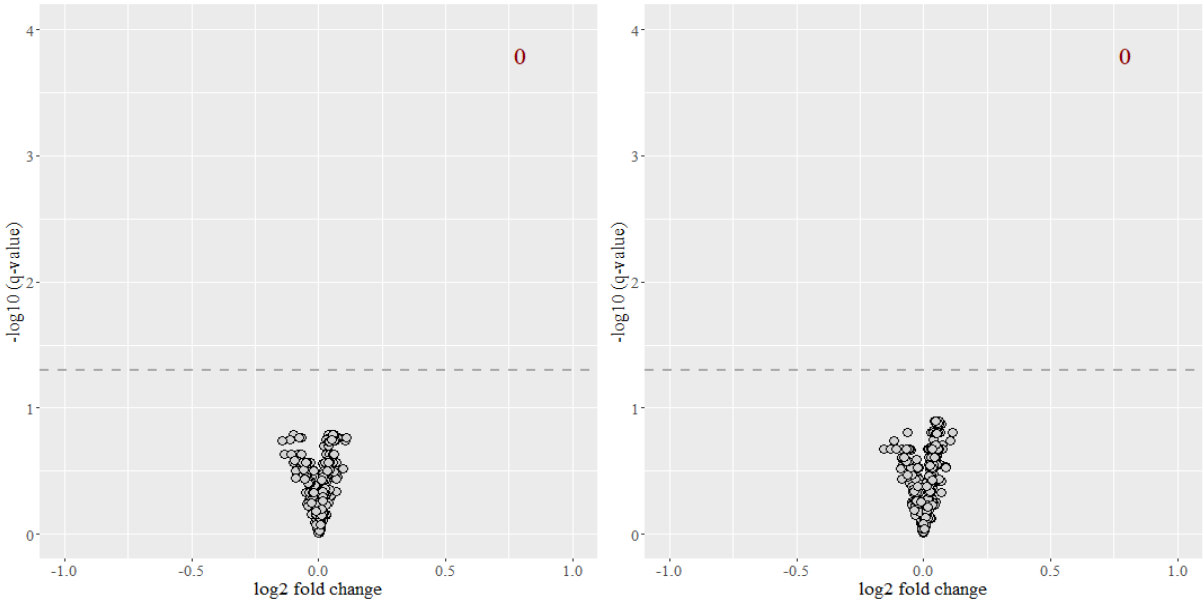


Fig. 84: Volcano plots for the comparison of Stage I and Stage III for C (left) and M (right).

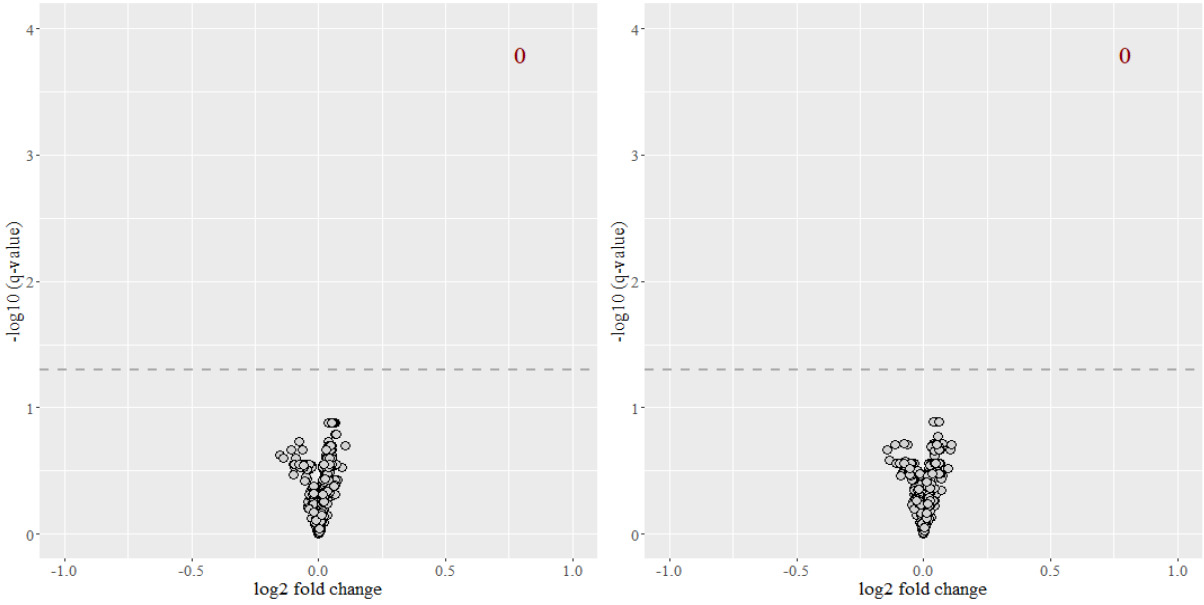


Fig. 85: Volcano plots for the comparison of Stage I and Stage III for T (left) and the combination of CM (right).

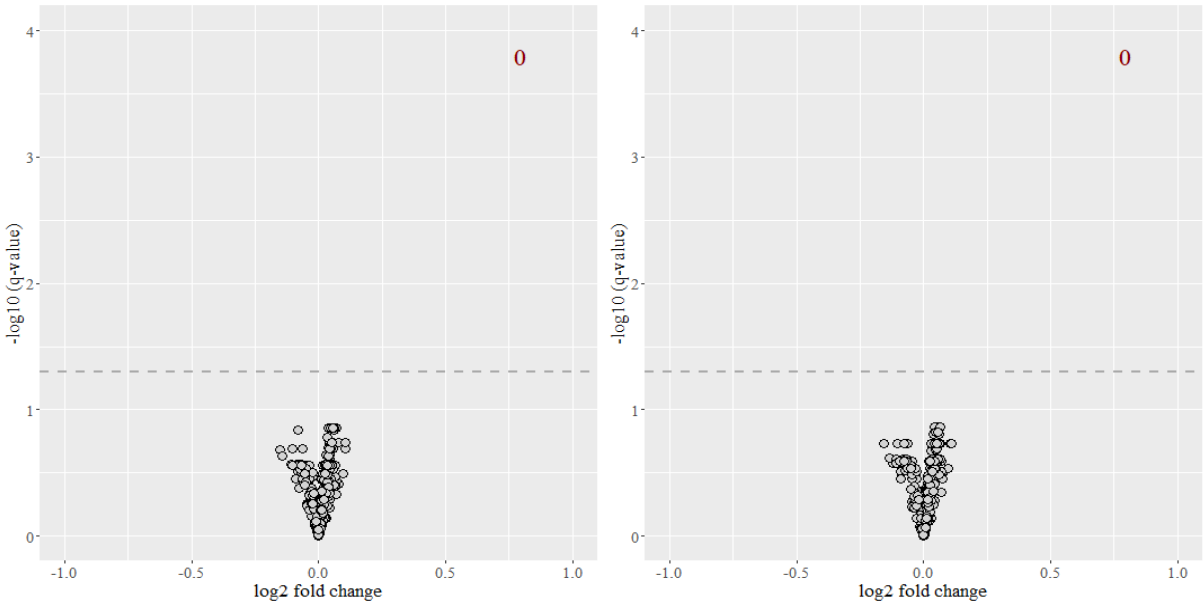


Fig. 86: Volcano plots for the comparison of Stage I and Stage III for the combination of CT (left) and the combination of MT (right).

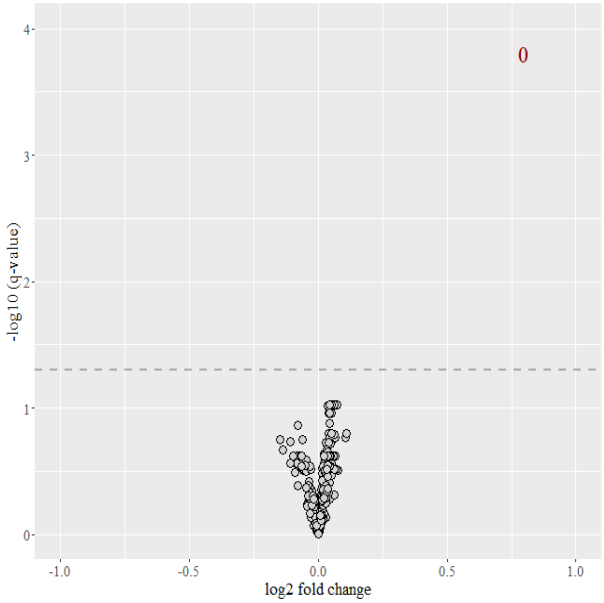


Fig. 87: Volcano plot for the comparison of Stage I and Stage III for the combination of CMT .

14.3.3 Stage I vs. Stage IV

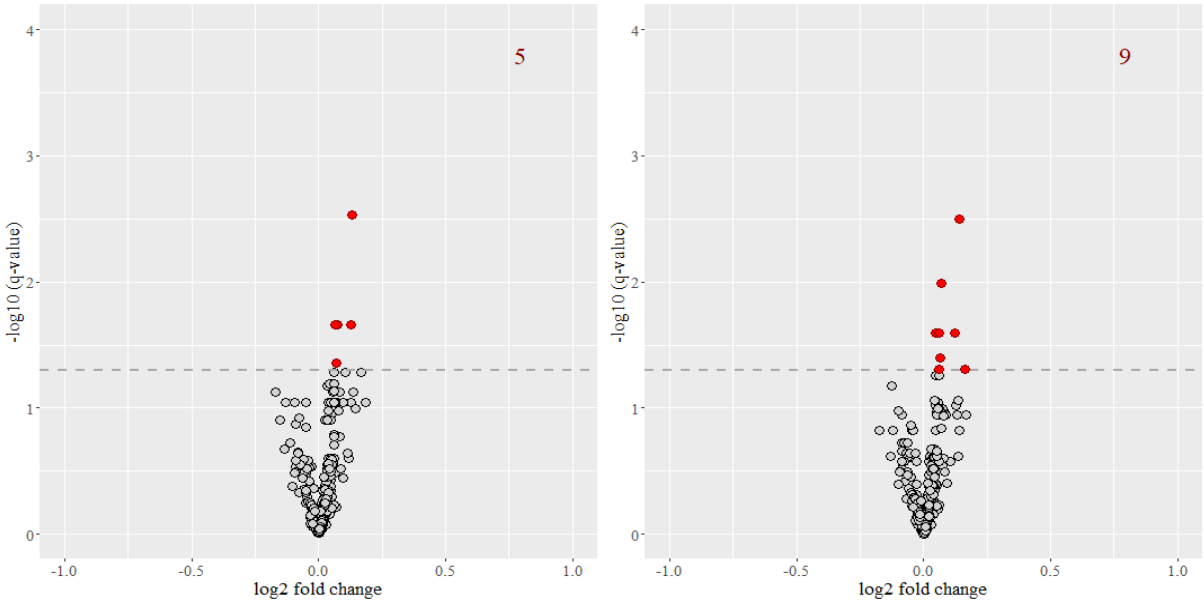


Fig. 88: Volcano plots for the comparison of Stage I and Stage IV for C (left) and M (right).

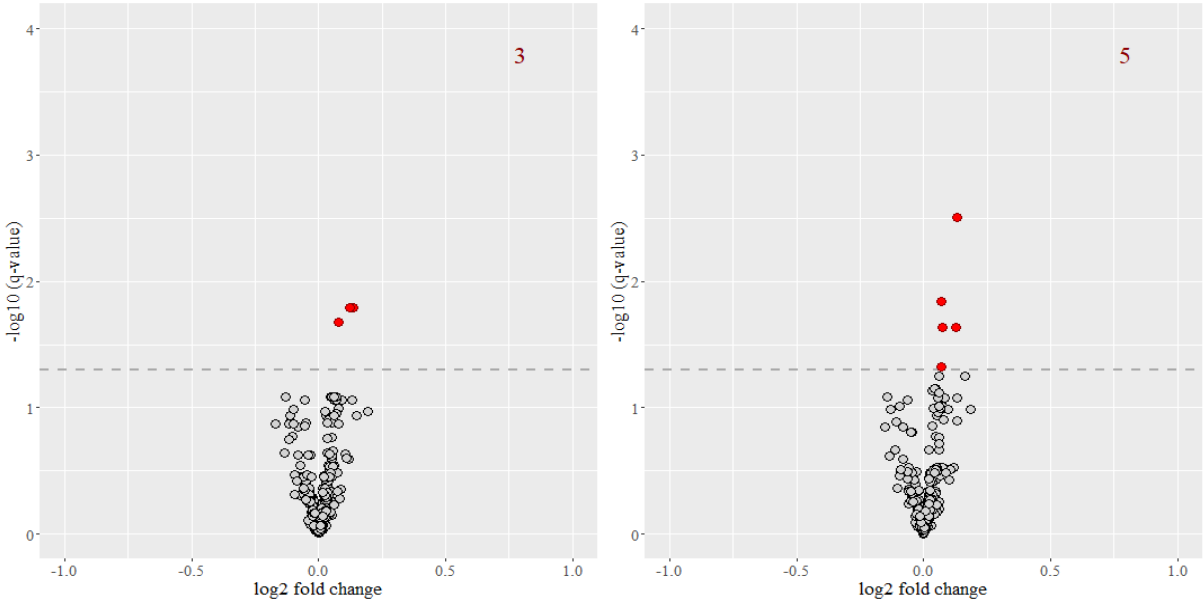


Fig. 89: Volcano plots for the comparison of Stage I and Stage IV for T (left) and the combination of CM (right).

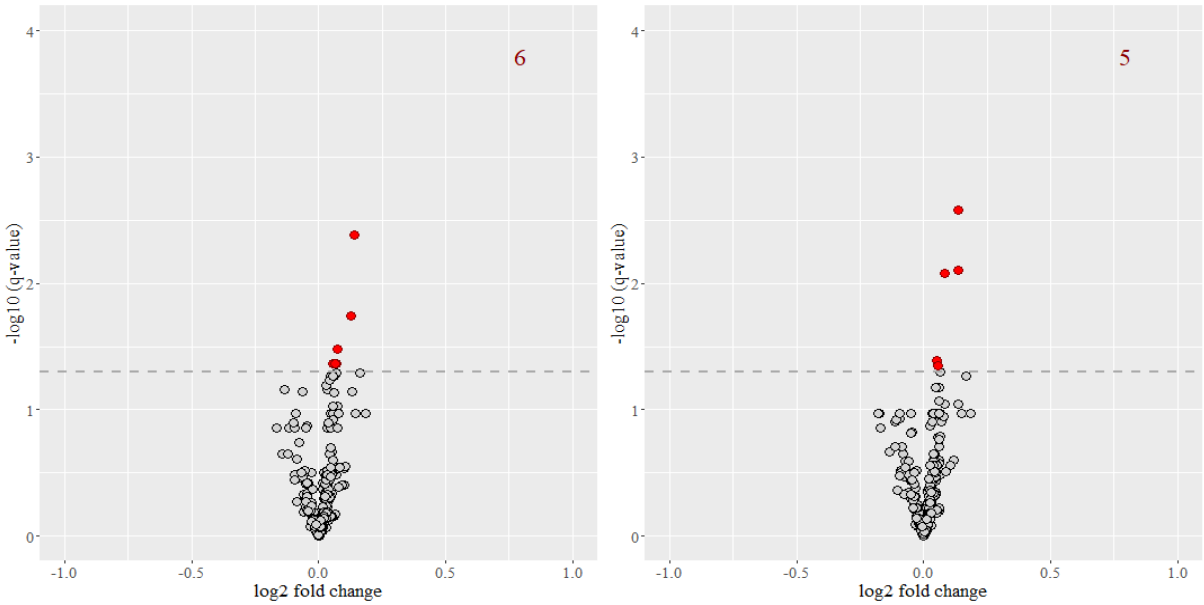


Fig. 90: Volcano plots for the comparison of Stage I and Stage IV for the combination of CT (left) and the combination of MT (right).

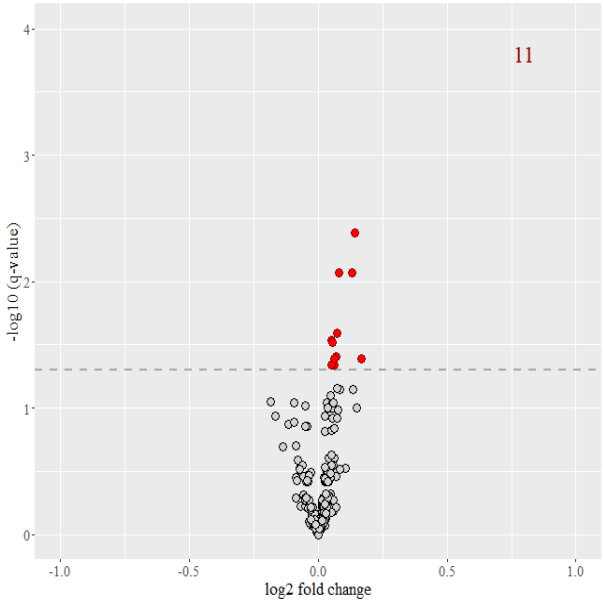


Fig. 91: Volcano plot for the comparison of Stage I and Stage IV for the combination of CMT .

14.3.4 Stage II vs. Stage III

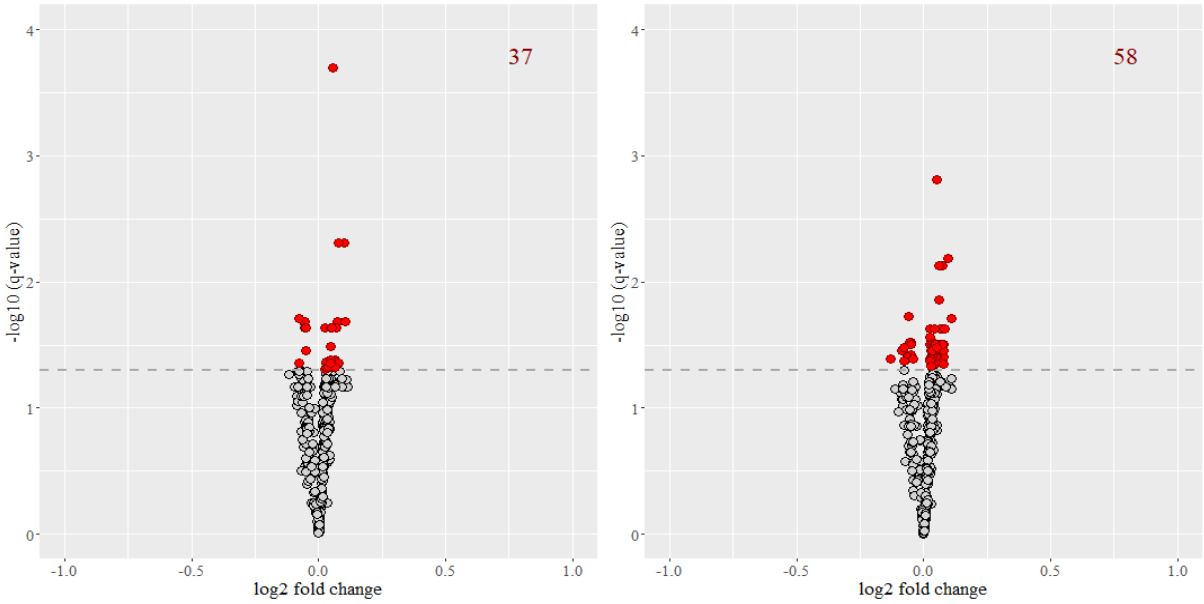


Fig. 92: Volcano plots for the comparison of Stage II and Stage III for C (left) and M (right).

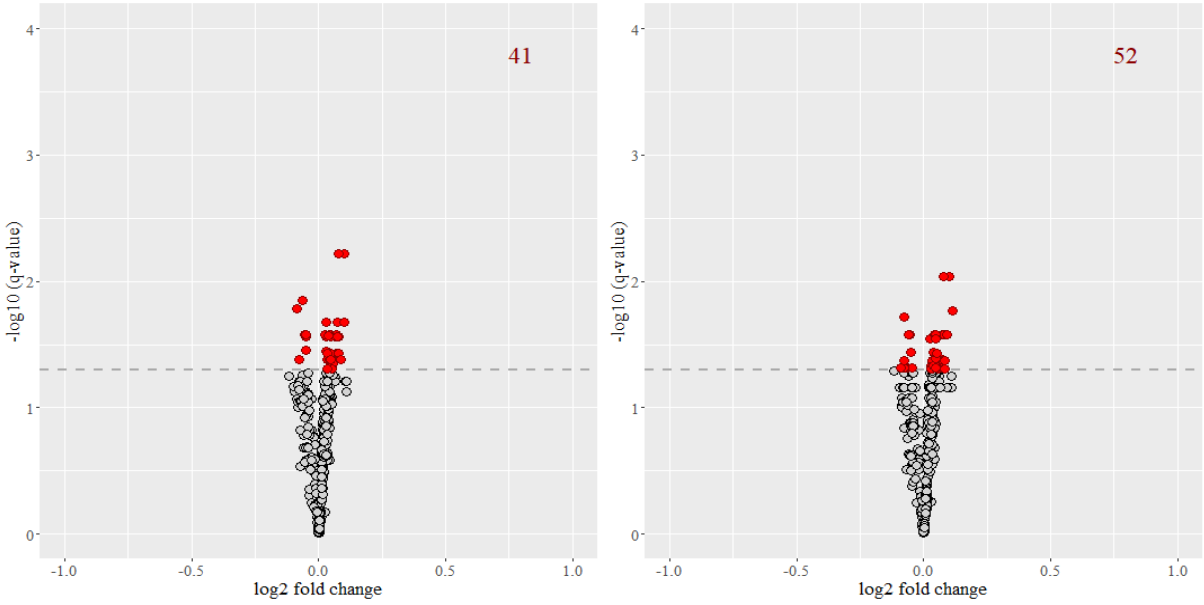


Fig. 93: Volcano plots for the comparison of Stage II and Stage III for T (left) and the combination of CM (right).

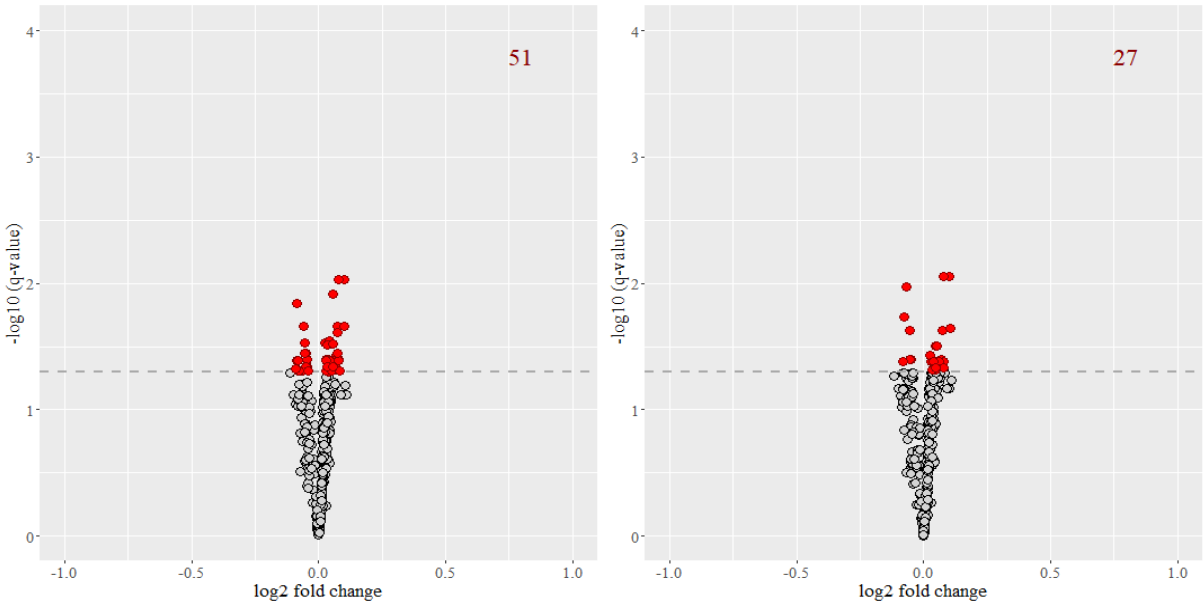


Fig. 94: Volcano plots for the comparison of Stage II and Stage III for the combination of CT (left) and the combination of MT (right).

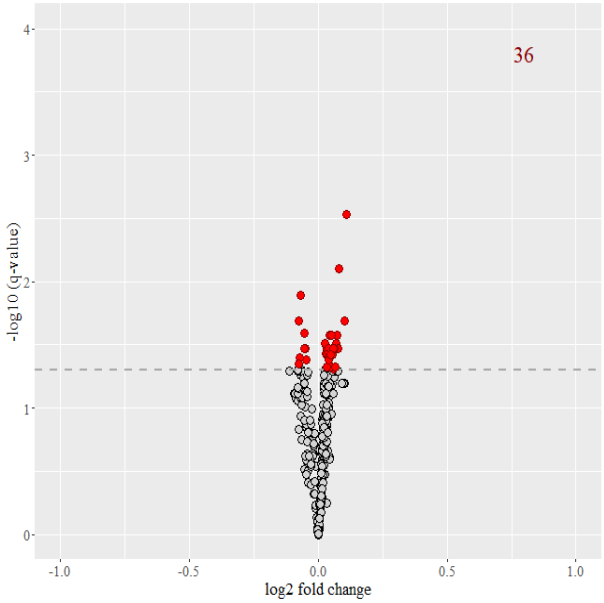


Fig. 95: Volcano plot for the comparison of Stage II and Stage III for the combination of CMT .

14.3.5 Stage II vs. Stage IV

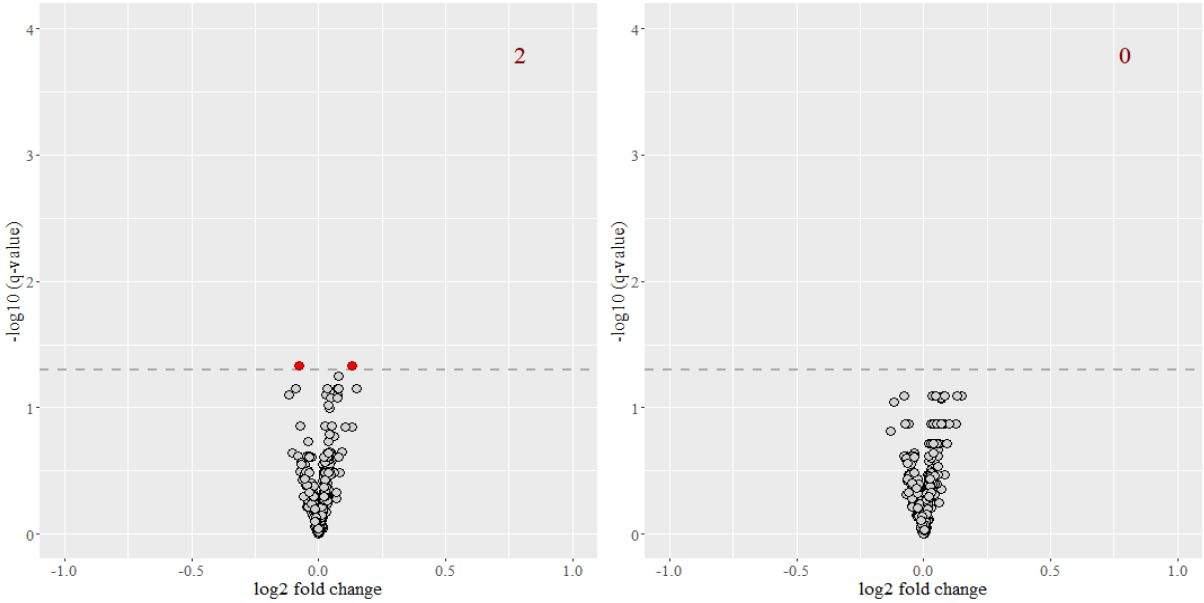


Fig. 96: Volcano plots for the comparison of Stage II and Stage IV for C (left) and M (right).

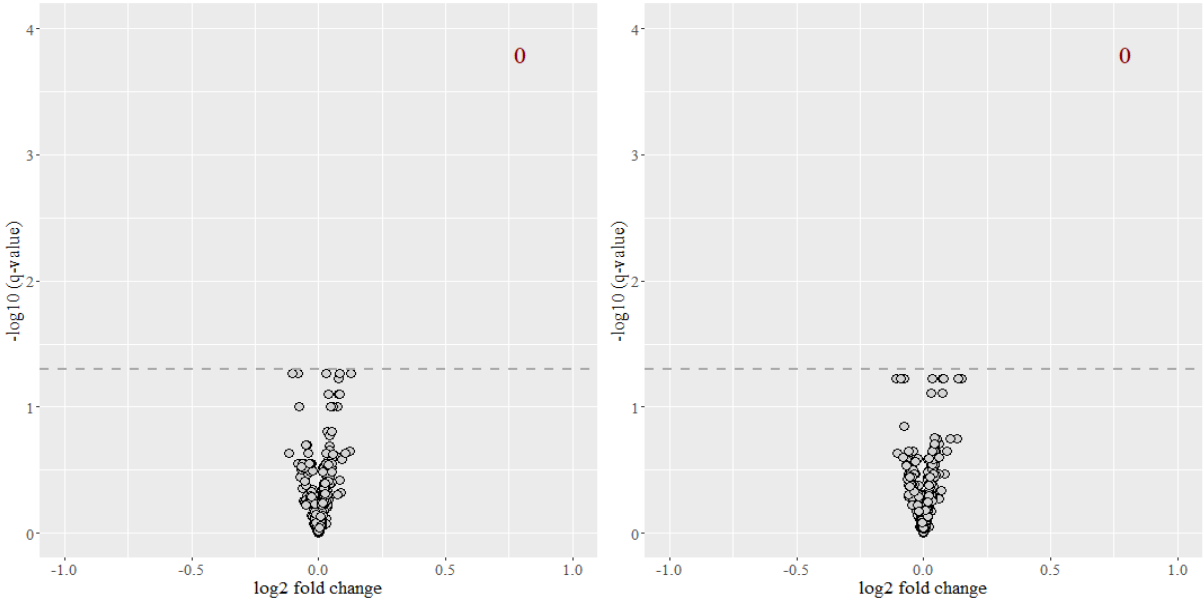


Fig. 97: Volcano plots for the comparison of Stage II and Stage IV for T (left) and the combination of CM (right).

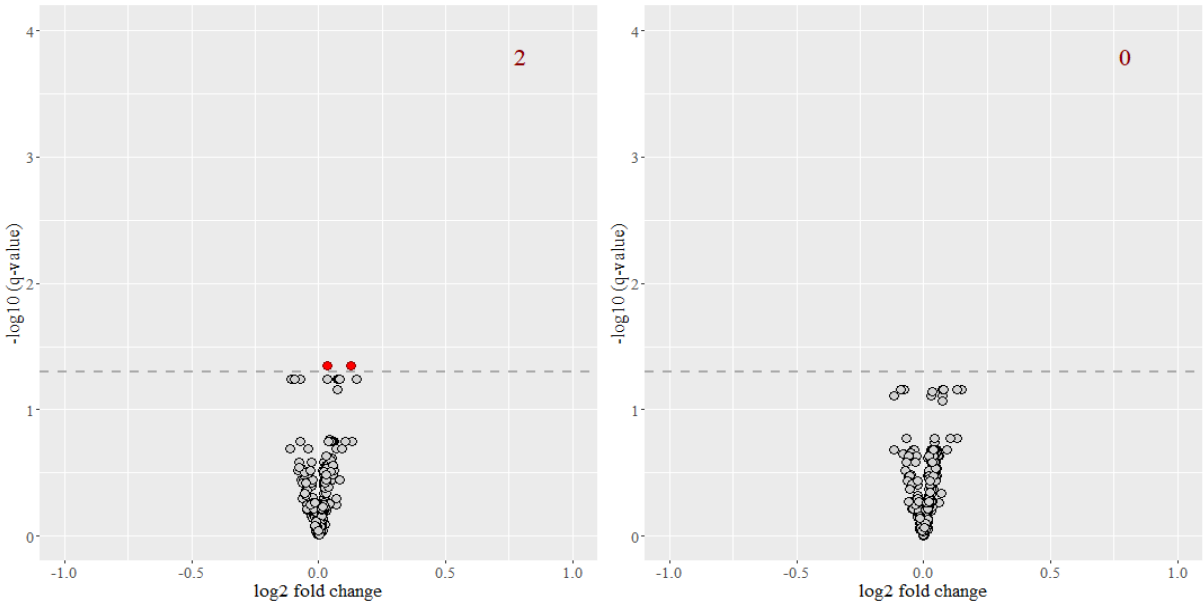


Fig. 98: Volcano plots for the comparison of Stage II and Stage IV for the combination of CT (left) and the combination of MT (right).

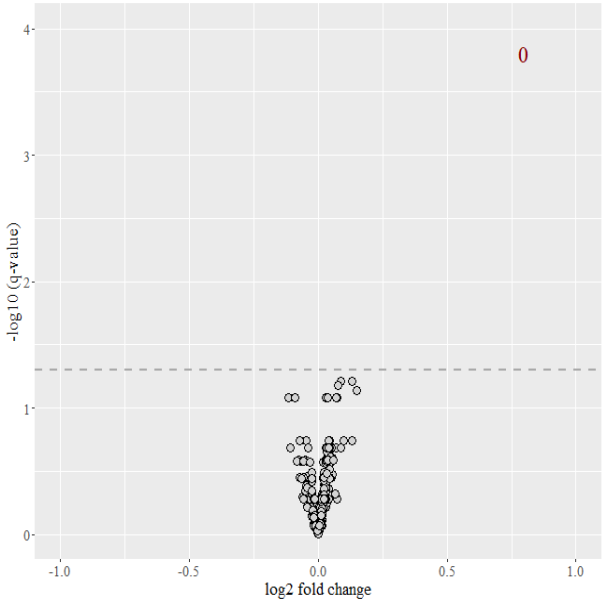


Fig. 99: Volcano plot for the comparison of Stage II and Stage IV for the combination of CMT .

14.3.6 Stage III vs. Stage IV

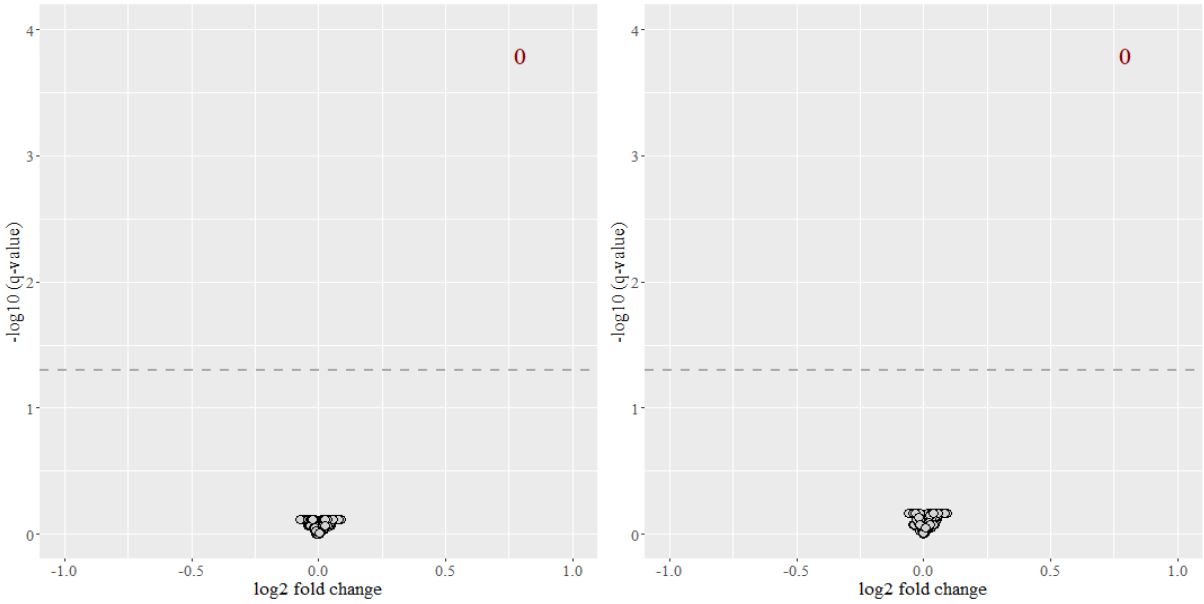


Fig. 100: Volcano plots for the comparison of Stage III and Stage IV for C (left) and M (right).

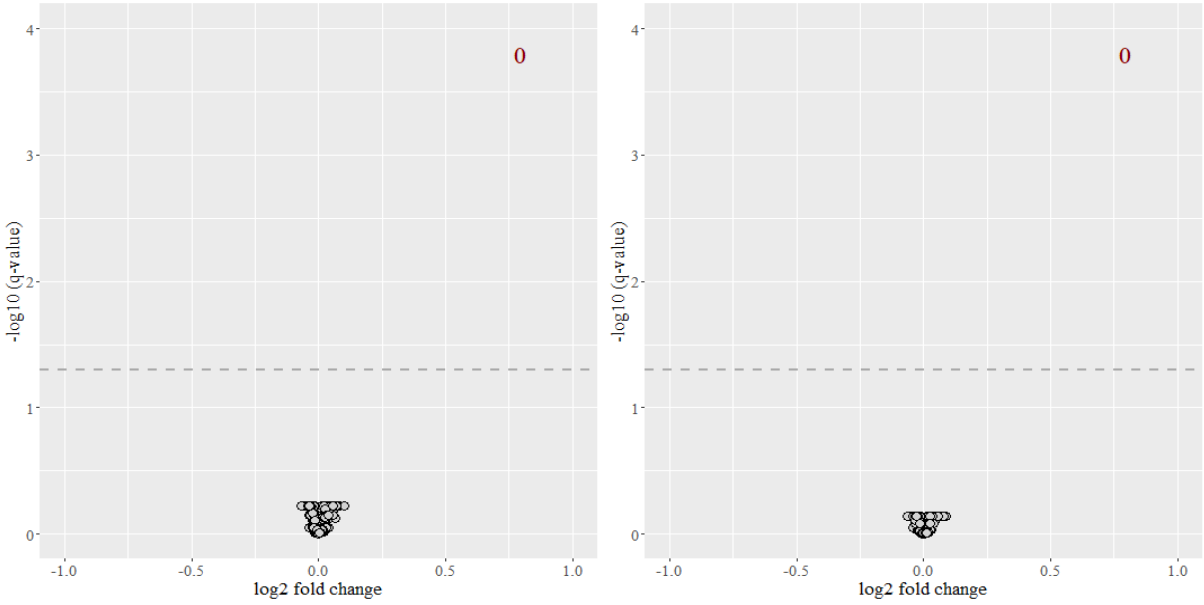


Fig. 101: Volcano plots for the comparison of Stage III and Stage IV for T (left) and the combination of CM (right).

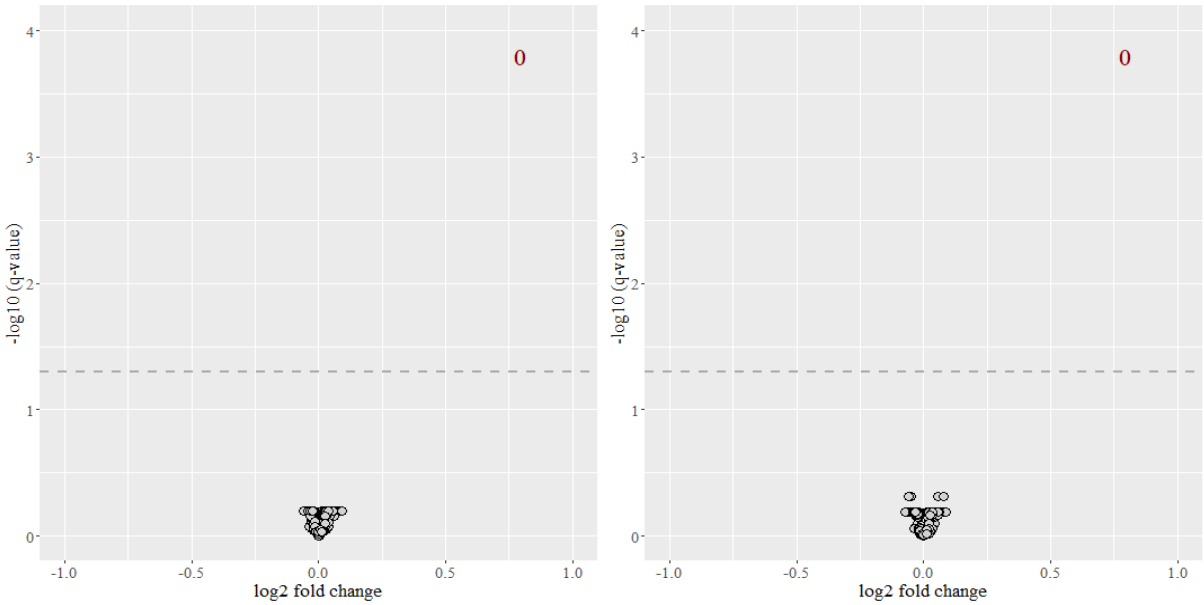


Fig. 102: Volcano plots for the comparison of Stage III and Stage IV for the combination of CT (left) and the combination of MT (right).

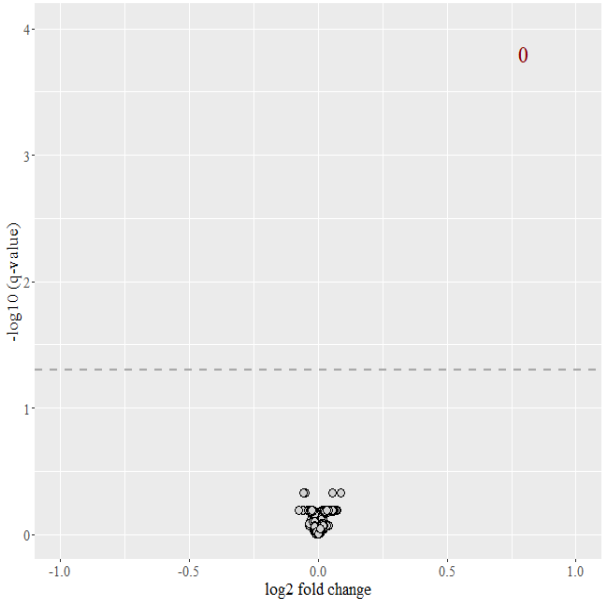


Fig. 103: Volcano plot for the comparison of Stage III and Stage IV for the combination of CMT .

15. Acknowledgements

First, I would like to thank my supervisor Prof. Dr. Hartmut Schlüter for welcoming me into his research group and for giving me freedom regarding the focus and realization of my project.

Second, I would also like to thank Prof. Dr. Chris Meier for accepting to be second supervisor and reviewer of the thesis.

Moreover, I would like to thank all my colleagues for the pleasant work atmosphere; especially I would like to express my gratitude to Dr. Christoph Krisp for his sound advice and for proofreading my thesis.

In addition, I would like to thank Hannah, Martina, and Kilian for the effort and time they put in to proofread the thesis. Thank you!

Außerdem geht ein riesiger Dank an meine Familie, die mich über die Jahre immer unterstützt hat; insbesondere an meine Eltern, die mir die Freiheit ermöglicht haben und das Vertrauen geschenkt haben, meine Ziele zu verfolgen.

Besonderer Dank gilt Tina.

16. Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, die vorliegende Dissertation selbst verfasst und keine anderen als die angegebenen Hilfsmittel benutzt zu haben. Die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium. Ich versichere, dass diese Dissertation nicht in einem früheren Promotionsverfahren eingereicht wurde.

……..

Hamburg, 11.08.2020